Revealing the Evolutionary History and Epidemiological Dynamics of Emerging RNA viral pathogens

Jayna Raghwani

Doctor of Philosophy
Institute of Evolutionary Biology
School of Biological Sciences
University of Edinburgh
2011

Abstract

Fast-evolving RNA viruses are a leading cause of morbidity and mortality among human and animal populations, contributing significantly to both global health and economic burden. The advent and revolution of high-throughput sequencing has empowered phylogenetic analyses with increasing amounts of temporally and spatially sampled viral data. Moreover, the parallel advancement in molecular evolution and phylogenetic methods has provided investigators with a unique opportunity to gain detailed insight into the evolutionary and epidemiological dynamics of emerging viral pathogens.

Using state-of-the-art statistical approaches, this thesis addresses some of the important but controversial questions in viral emergence. Chapter 2 introduces a new framework to quantify and investigate reassortment events in influenza A viruses. By developing a computationally efficient algorithm to calculate the largest common subtree for a pair of tree sets, which are estimated from different parts of the genome for the same taxa set, the level of phylogenetic incongruency due to reassortment can be appropriately ascertained. Chapters 3, 4 and 5 investigate the evolutionary origins of three different viruses: the novel emergence and cross-species transmission of SARS-CoV, the genesis and dissemination of the unique HCV circulating recombinant form, and the ancient divergence of all influenza viruses, respectively. Moreover, Chapter 4 presents an improved statistical framework, which provides more precise evolutionary estimates, by utilizing the hierarchical bayes approach to investigate recombination events in emerging RNA viruses. The last empirical study, presented in Chapter 6, applies the recently developed Bayesian phylogeography models to a large viral sequence dataset sampled from southern Viet Nam to examine the fine-scale spatiotemporal dynamics of endemic dengue in Southeast Asia.

The work presented here reflects both the advancements made in sequencing technology and statistical phylogenetics, along with some of the challenges that remain in studying the emergence of fast-evolving RNA viruses. This thesis proposes new and improved solutions to these evolutionary problems, such as incorporating non-vertical evolution (i.e. homologous recombination and reassortment) into the phylodynamic framework, with the aim of facilitating future investigations of emerging viral diseases.

Acknowledgements

The worked presented in this thesis would not have been possible without the guidance, support and friendship of several individuals. First and foremost, I would like to thank my supervisor Andrew Rambaut for introducing to me the world of phylogenetics and infectious diseases. I am grateful for his enthusiasm and insight throughout my PhD. I have greatly appreciated his patience and good nature as a supervisor, especially when dealing with my programming woes. As a complete neophyte to computational biology, I owe special thanks to Sam Lycett, for consistently making time to discuss my queries about statistics, algorithms and java.

Throughout my PhD, I have also had the opportunity work with collaborators outside of Edinburgh. For the dengue project, I would like to thank Cam Simmons and colleagues in Ho Chi Minh City and the Broad Institute for providing the viral sequence data. Additionally, I am grateful to have had the chance to collaborate with Eddie Holmes, who along with Andrew and Cam, contributed to many enlightening discussions on this project. For the Hepatitis C Virus project, I am greatly indebted to Oliver Pybus for his supervision and scientific diligence. I would also like to thank Xiomara Thomas, Sylvie Koekkoek, Janke Schinkel, Richard Molenkamp and Thijs van de Laar for providing new sequence data and imparting their virology expertise. In addition, I have appreciated the helpful interactions with Yutaka Takabe, Yasuhito Tanaka, and Masashi Mizokami on the history and epidemiology of HCV. At Edinburgh, I would like to thank the members of Virus club and the wider community of evolutionary biologists in Ashworth. Their academic discourse and advice over the years has been an invaluable source as a fledgling scientist.

This process has been made easier with the moral support of some wonderful friends. Besides supplying much needed chocolate and cake to fuel the scientific process, I would like to thank the following people for providing diversions during my PhD. In no particular order, they are: Claire Webster, Claire Bourke, Jess Hedge, Steph Smith, Laura Pollitt, Adam Hayward, Ricardo Ramiro, Gethin Evans, Fiona Lethbridge, Harriet Stone, Heidi Kuehne, Michelle Clements, Laura Appleby, Matty Hartfield, Jen Scholefield, Carolyn Riddell, Amy Jennings, Cheryl Gibbons, Richard Perry, Sinead Collins, Vera Kaiser, Beatriz Vicoso, Penny Haddrill, Lel Eory, Paula MacGregor, Cat Prendegast, Ewa Jurneczko, Magnus Kelly, Lara Salido, Nisha Hirani and Roshni Shah. A special thanks goes to Laura Emery and Melissa Ward, for being both great colleagues and awesome friends. Finally, I would like to thank my family. Their good humour, patience and sage advice have kept me grounded. And for their complete support and love over the years, I am eternally grateful.

Declaration

I declare that I composed this thesis myself, and that the work described herein is my own except where explicitly stated below. This work has not been submitted for any degree or professional qualification except as specified.

Jayna Raghwani, 2012

Contents

1

General Introduction

1.1 Emerging Viruses

Viruses are very adept subcellular parasites that pervade all domains of life (Prangishvili et al., 2006). The great diversity observed in their biology indicates a complex evolutionary history, perhaps reflecting that multiple independent events have been involved in their origins. The notorious lack of fossil record for viruses has meant that investigating their early emergence has proven to be challenging. However, it is clear that viruses have ancient origins, but whether it preceded the diversification of life or arose shortly after remains to be determined (Koonin et al., 2006; Iyer et al., 2006). This observation is supported by the widespread presence of the antiviral system in eukaryotes, the RNA interference (RNAi) pathway (Cerutti and Casas-Mollano, 2006).

The dependency on other living beings means that these small infectious agents have coevolved with their hosts, resulting in a long-standing evolutionary arms race, that has shaped both their genomes and ecology. Viral genomes show marked variation in size, content and organization, encoding the genetic material either in RNA or DNA. This in part reflects the enormous evolutionary capacity of the obligate parasites, which can differ considerably in their adopted lifestyles, to adapt to a broad range of environments.

The term emerging viruses has been used to describe viral pathogens that have arisen from a zoonotic (animal-infecting) source. They may have infected the host since the last common ancestor, with sister host species harbouring closely related viruses, e.g. GBV-C and hepatitis B viruses (Simmonds, 2001; Fares and Holmes, 2002). Alternatively, they may represent a completely new cross-species transmission where the virus has jumped from an ecologically linked species, either through predation or by occupying similar niches. Emerging viruses are responsible for a large proportion of infectious diseases in the natural population, and can have a great impact on the biodiversity. A notable example is the appearance of SIV in chimpanzees from Old World monkeys in Africa, which has attributed to their dramatic population decline (Rudicell et al., 2010).

Fast-evolving RNA viruses, compared to their DNA counterparts, are particularly

overrepresented amongst emerging infectious diseases (Woolhouse, 2002; Duffy et al., 2008). The extremely high evolutionary rates, short generation times and large population sizes, enable RNA viruses to be exceptionally successful pathogens by overcoming cross-species barriers and maintain ongoing infections. While some DNA viruses can display evolutionary rates on similar scales, e.g. canine parvoviruses (Shackelton et al., 2005), their general characteristics as pathogens are associated with persistent infections and vertical transmission (Duffy et al., 2008; Villarreal et al., 2000; Holmes, 2003b, 2008). Due to the considerable risk posed to global health and economy, this thesis focuses on RNA viruses that have recently emerged in the human population.

1.1.1 Viral Emergence

There are a number of steps involved in viral emergence, demonstrating the evolutionary and ecological challenges that viruses need to overcome to infect new host species (Morse, 1995; Holmes, 2006; Wolfe et al., 2007). First the virus must be able to enter the host cells, via their cell surface receptors. Once inside the cell, the virus needs to replicate efficiently to produce enough virus particles for the next infection. Lastly, to fully emerge into the new species, the virus needs to establish host-to-host transmission and spread effectively in the population (Cleaveland et al., 2001; Woolhouse et al., 2005; Wolfe et al., 2007).

Although viral outbreaks are often perceived as opportune moments from the pathogen perspective, the reality is that a series of factors increase the likelihood of a novel cross-species transmission (Holmes and Rambaut, 2004; Woolhouse et al., 2005). Evolutionary relationship and ecological networks of the host species are considered to be the major determinants of predicting potential viral reservoirs (Baranowski et al., 2001; Woolhouse, 2002). Closely related species are more likely to share similar biological features, such as cell surface receptors, immune system and host replication factors, which expectedly can enable viruses to jump more easily between hosts (Woolhouse, 2002). However overlapping habitats and frequency of exposure to new pathogens may have greater importance in determining the actual risk of emergence (Woolhouse et al.,

2005; Davies and Pedersen, 2008). For example, the most probable viral reservoirs for humans are rodents, birds and ungulates, compared to other primates, which reflects the relative probability of humans interacting with other primates are small (Cleaveland et al., 2001).

The last two steps in viral emergence concerns transmission, and are largely driven by within and between host factors. RNA viruses with their fast mutation rates, acute infectious periods, and thus large population sizes, can overwhelm the host immune system and be transmitted effectively through the population. The mode of transmission (how they are spread between individuals, e.g. by aerosol or bodily fluids), may have additional consequences on the viral epidemiology (Woolhouse et al., 2005). This latter point can be illustrated by contrasting the SARS and Ebola outbreaks (Woolhouse et al., 2005), where both emerged from bats (Leroy et al., 2005; Li et al., 2005). SARS rapidly disseminated to different countries, whereas Ebola was associated with a localised epidemic. The main differences between these viruses are: the location of emergence and how they are transmitted in the population (Woolhouse et al., 2005). SARS is an airborne pathogen that emerged in one of the most densely populated regions of the world, southeast China. In contrast, transmitted by bodily fluids, Ebola has largely appeared in remote parts of Africa. Another distinction is observed between the infection periods of these viruses, with Ebola virus showing a more variable generation time (1-21 days)(Francesconi et al., 2003) compared to SARS (2-7 days). Therefore, to understand the disease dynamics of emerging RNA viruses, we need to understand factors that affect both the virus evolution and the host population.

1.1.2 Human Ecology and RNA Viruses

A major implication of acute infecting RNA virus ecology is that they are normally maintained by large, densely populated and social animal reservoirs (Dobson and Carper, 1996; Diamond, 2002). Thus, these pathogens are only likely to have become a threat to humans when agricultural societies started to develop, which paved the way for larger and better connected communities, along with close interactions with animals

(Cleaveland et al., 2001, 2007; Diamond, 2002). The earlier hunter-gatherer humans almost certainly encountered RNA viruses, either from prey or other wildlife sources, but due to their small groups they are unlikely to have sustained long transmission chains for RNA viruses to persist in the population. The current repertoire of RNA viruses in humans have further benefited from escalated population expansion and increasingly crowded habitats (Pearce-Duvet, 2006; Diamond, 2002). This has largely been fuelled by technological advancements of the industrial revolution, which mechanised many human activities, such as rural farming and travel, significantly transforming the way we live today. In turn, urbanisation and human movement have become major drivers of emerging viral diseases, facilitating endemic transmission in populations and promoting rapid global dissemination (Cleaveland et al., 2001, 2007).

Furthermore, the accompanying anthropogenic modifications of the environment have had significant repercussions on the transmission ecology RNA viruses that affect humans (Pearce-Duvet, 2006). Deforestation of wildlife habitats has brought us in closer proximity to unknown zoonotic sources (Wolfe et al., 2005), while construction of transport networks (railways and roads) have made these areas more accessible (Cleaveland et al., 2001, 2007; Schrag and Wiener, 1995). These factors can explain the relative recent origins of HIV, dating around the 1930s (Korber et al., 2000), even though it is evident that humans in Africa have had frequent and older exposure to the virus prior to this time (Worobey et al., 2008). Humans have also transformed the way zoonotic pathogens are transmitted in the natural world, which is illustrated by recent examples of human-mediated spread of wildlife rabies in North Africa and transmission of swine influenza A viruses (Talbi et al., 2010; Nelson et al., 2011). The process of viral emergence in humans often involves indirect transmission pathways from the original zoonotic source, such that multiple hosts are implicated in the cross-species transmission (Cleaveland et al., 2001). These secondary hosts are formally referred to as intermediate or incidental reservoirs, denoting their roles in either propagating or facilitating transmission to different species, rather than being the original or the main viral source. Intermediate hosts can assist viruses to circumvent or adjust the steep

fitness valleys that they may need to cross to competently emerge into alternate hosts (Kuiken et al., 2006). In the case of human influenza A viruses, pigs are considered to have significant roles due to their possession of both human-like and avian-like receptors, thus allowing avian viruses to emerge in humans more readily (Scholtissek, 1987; Kida et al., 1994; Webster et al., 1992). In addition, some incidental hosts can serve to amplify the prevalence of the viral pathogen in the natural environment, thereby increasing the transmission likelihood to humans (Pearce-Duvet, 2006). This has been observed in human outbreaks that originated from bat reservoirs, for example Hendra, Nipah and Ebola viruses, where rodents, swine and horses have played a crucial part in viral propagation (Chua et al., 1999; Mackenzie, 1999; WHO, 2008).

1.2 The Phylodynamic Framework

The phylogenetic method has become an invaluable tool in the field of molecular epidemiology, especially with respect to fast evolving pathogens like RNA viruses (Pybus et al., 2001; Drummond et al., 2003a; Grenfell et al., 2004). The continued advancements in both sequencing and computer technologies have enabled phylogenetic analyses with both great power and proficiency to study large data sets of viral samples (Pybus and Rambaut, 2009). Furthermore the marriage of epidemiology and phylogenetics into a unified framework, which is called phylodynamics, has revolutionized evolutionary and ecological investigations of RNA viruses (Grenfell et al., 2004). The central concept behind the phylodynamic framework is that epidemiological processes, at the intra- and inter- host levels, shape the viral genetic diversity (Grenfell et al., 2004). This is mostly down to the observation that the timescales of the evolutionary change and the population dynamics of RNA viruses are effectively the same, due to their extremely high mutation rates and large population sizes. (Drummond et al., 2003a; Grenfell et al., 2004; Pybus and Rambaut, 2009). Therefore, the phylogenetic reconstruction of serially sampled viral sequences can reveal not only the underlying evolutionary relationship, but also the ecological and population processes shaping their immediate histories (Rodrigo and Felsenstein, 1999; Pybus et al., 2000; Drummond

et al., 2003a; Grenfell et al., 2004)

1.2.1 Background

The coalescent is a crucial component to the phylodynamic framework to infer the population dynamics and substitution rates of fast-evolving RNA viruses (Rodrigo and Felsenstein, 1999; Pybus et al., 2000; Drummond et al., 2003a; Grenfell et al., 2004). The coalescent is a model to trace the ancestry of randomly selected individuals from a neutrally evolving population until there is only a single common ancestor (Kingman, 1982). The population history or gene genealogy is inferred backwards in time. Starting with the sampled taxa (and their corresponding gene sequences), which represents the extant lineages in the population, the coalescent theory estimates when these lineages share a common ancestor through time. The process of finding the common ancestor of a pair of lineages is called coalescence, such that two lineages become one. Eventually in this model all lineages coalesce, which is indicated by the most recent common ancestor of the taxa set. The timing of the coalescent event or probability of sharing a common ancestor in the genealogy is a function of the population demographic history (Kingman, 1982). Different population processes, such as exponential growth or constant size, are associated with distinct patterns of coalescence and genetic diversity of the sampled sequences (Rodrigo and Felsenstein, 1999; Pybus et al., 2000; Drummond et al., 2003a; Grenfell et al., 2004). This observation is key to the phylodynamic framework, since we can infer the underlying population dynamics directly from the viral sequences. The introduction of non-parametric coalescent growth models, such as the skyline (Strimmer and Pybus, 2001; Pybus et al., 2000; Drummond et al., 2005) and more recently the skyride (Minin et al., 2008), have further increased the applicability of the coalescent to sequence data, by accommodating more complex patterns of demographic history. The earlier classical skyline plot works on the principle that effective population size can change at coalescent events, but remain constant during the coalescent interval (time between coalescent events). This method have been subsequently extended to address the noisy demographic parameter estimates, e.g generalized skyline plot (Strimmer and

Pybus, 2001), and the time structure of the sequences, e.g. Bayesian skyline and GMRF skyride plots (Drummond et al., 2005; Minin et al., 2008). The latter has been enabled by the extension of the coalescent to consider heterochronous sampled sequences, thus allowing the explicit consideration of time into coalescent-based inferences (Rodrigo and Felsenstein, 1999). This has been found to be especially fruitful for studying fast-evolving RNA viruses, where the processes of interest, such as substitution rates and population size, may change through time (Rodrigo and Felsenstein, 1999; Pybus et al., 2000; Drummond et al., 2003a). Thus by utilizing the temporal structure of the sampled sequences, the increased statistical power has greatly improved the precision of population demographic and evolutionary estimates (Drummond et al., 2003a).

The molecular clock hypothesis assumes that the rate of evolution along a lineage is a function of time. It is an extremely powerful tool for disentangling the number of substitutions and time along the branch in the phylogeny, allowing evolutionary biologists to estimate evolutionary rates from sequence data and infer divergence times in the phylogeny. In order to do this, the molecular clock needs to be calibrated independently with some time information. For fast-evolving sequences like RNA viruses that have been sampled over time, the calibration points are typically the sample dates of the viral isolates. For contemporaneously sampled sequences, the molecular clock can be calibrated by assigning a date or a time range to an internal node in the phylogeny. In this case the information typically comes an external source, such as a historical or fossil data. The practicality of the molecular clock not only allows us to estimate evolutionary rates and divergence times from sequence data, but also infer the root position on the phylogeny without defining an outgroup *a priori*. The strong assumptions of the molecular clock hypothesis, suh that the rate is either constant (strict) or unconstrained (independent rates for each branch) throughout tree have often been challenged to be biologically unrealistic or too restrictive (Drummond et al., 2006; Thorne et al., 1998). At least for fast-evolving RNA viruses, the assumption of the strict molecular clock has been found to not hold particularly well in general, most likely due to variation in the viral generation times (Jenkins et al., 2002; Drummond et al., 2006). The introduction

of the relaxed molecular clocks has provided a practical alternative to the two extreme models of the molecular clock hypothesis (Drummond et al., 2006). Different strategies have been taken to model the rate variation throughout the tree, ranging from local clocks (Yoder and Yang, 2000; Drummond and Suchard, 2010), to autocorrelated (Thorne et al., 1998; Drummond et al., 2006) and uncorrelated clocks (Drummond et al., 2006). A local molecular clock allows different parts (typically predefined by the investigator) of the tree to evolve at different evolutionary rates, which may be appropriate for studying taxa with significantly different life histories or tempo of evolution. Autocorrelated clocks work on the principle that the rates of the descendant lineages in a tree are a function of the parental branches (Thorne et al., 1998). This group of clock models assume that the rate variation among branches is predominantly explained by inherited factors, such as generation time or metabolic rates. With uncorrelated clocks, the rates of each branch in the tree is independently drawn from a parametric distribution, such as a log-normal or an exponential distribution (Drummond et al., 2006). The basis of uncorrelated clocks are that the among-branch rate variation is mostly driven by non-inheritable or stochastic factors. For viral sequences that have been sampled over relatively small timescales, e.g. ten to hundreds of years, the best fitting clock models to explain the rate variation along the tree have been found to be the uncorrelated clocks (Drummond et al., 2006).

The phylodynamic framework has also greatly benefited from the advancements of statistical phylogenetics, namely by incorporating maximum-likehood and Bayesian phylogenetic inference (Felsenstein, 1981, 1973; Larget and Simon, 1999; Mau et al., 1999; Mau and Newton, 1997; Yang and Rannala, 1997). The main appeal of both these approaches over the traditional parsimony and distance-based methods (e.g. neighbour-joining phylogenetic reconstruction) is that they provide some measurement of confidence about the evolutionary models applied, such as the substitutional or the coalescent model, to the sequence data to estimate the phylogeny. Furthermore, introducing a statistical framework into the phylogenetic estimation enables to directly test different evolutionary models on the sequence data. In maximum-likelihood phylogenetics, the

aim is to find the most probable tree given the sequence data and chosen evolutionary model. This requires an initial guide tree, estimated either by neighbour-joining or parsimony method, from which the model parameters are estimated. Subsequently, a heuristic search algorithm is employed to find the maximum-likelihood phylogeny based on tree-pruning algorithms to explore the tree space (i.e. the number of potential trees that could explain the observed sequences). As the sequence data becomes larger, the search for the most likely tree becomes increasingly difficult computationally as the tree space grows exponentially. Bayesian phylogenetic inference in contrast does not attempt to find the most likely tree, but instead estimates a distribution of plausible phylogenies that best explain the observed sequence data. Importantly, the tree search space is restricted by the prior information from the chosen evolutionary models. Thus, together with the likelihood of observing the data under the chosen models, the posterior tree distribution is calculated. This set of posterior trees can be used to determine the probability, i.e. the frequency of observing an event from all possible events, of a certain phylogeny. Specifically, this is known as the posterior probability, which can be inferred for any branching event observed in the posterior tree distribution. However, as with maximum-likelihood methods, estimating the posterior distribution analytically is computationally intractable for all but simple analyses, i.e. for small datasets with uncomplicated models. Fortunately, the introduction of the Markov Chain Monte Carlo (MCMC) technique into the Bayesian inference framework has alleviated this computational burden, which has enabled faster estimation of statistical support than maximum-likelihood methods (Larget and Simon, 1999; Mau and Newton, 1997; Mau et al., 1999; Yang and Rannala, 1997; Holder and Lewis, 2003). The MCMC is a sampling procedure to estimate the posterior tree distribution, where trees are sampled in proportion to their posterior probabilities. From a Bayesian perspective, the most likely tree will be sampled more frequently than a less probable tree.

An important aspect of Bayesian phylogenetic analysis is the prior choice and specification of the models employed to describe the sequence data (Holder and Lewis, 2003; Alfaro and Holder, 2006). Implementing different evolutionary models expresses dif-

ferent beliefs about how the data has been generated. The extent of the uncertainty (or certainty) we may have about these models can be specified by assigning a prior probability distribution for the model parameters (Alfaro and Holder, 2006). Commonly, these are known as simply 'priors', which can be generally be either informative or uninformative about the evolutionary process. When we have no real expectation of the parameter values, an uninformative or improper prior may be most appropriate for analysing the sequence data. These probability distributions have large variances, e.g. gamma or uniform distributions, indicating that a wide range of parameter values can potentially explain the observed data. In this case, the posterior estimates are informed by the sequence data itself, such that differences in the posterior probability of the MCMC sample will reflect the differences in the likelihood. Alternatively, informative priors may be employed if there is some known belief about the parameter space of the model. In most cases, largely uninformative priors are implemented, where some knowledge may be known about the lower and upper bounds, but are typically described by a large variance. A major concern in Bayesian inference is the effect of priors when estimating posterior probabilities for different hypotheses, which can lead to the problem of model misspecification or prior interaction (Holder and Lewis, 2003; Alfaro and Holder, 2006). Furthermore, due to the employment of complex evolutionary models in Bayesian phylogenetics, it is difficult to predict how these priors will behave. Therefore, preliminary data exploration with different priors is paramount before performing any Bayesian phylogenetic analysis. This can help inform about the prior choices and importantly gauge the suitability of the data for the question of interest (Alfaro and Holder, 2006).

In spite of these concerns surrounding prior choice in Bayesian analyses, phylogenetic packages that employ Bayesian MCMC have become popular over maximum-likelihood methods to investigate viral evolution and epidemiology. With the explosive growth in sequence data and parallel advances in computational techniques, the Bayesian methods have become powerful tools for evolutionary and molecular biologists. While the inception of the phylodynamic framework was initially formulated

with maximum-likelihood (Pybus et al., 2000, 2001; Drummond et al., 2003b), the Bayesian approach have become significantly favourable in light of these improvements in data and technology. The introduction of the MCMC algorithm has significantly reduced the time to estimate posterior probabilities for large datasets (Holder and Lewis, 2003). In contrast, the likelihood calculations are still comparatively slow in the maximum-likelihood phylogenetic framework. For the study of fast-evolving RNA viruses, utilizing the time of sampling into the phylogenetic analysis has made the Bayesian approach an especially powerful and intuitive framework. Besides providing interpretable units of evolution in the form of time-scaled phylogenies (Drummond et al., 2006), this has improved the precision of evolutionary estimates that are a function of time, such as substitution rates and population size dynamic (Drummond et al., 2003a). Furthermore, from the viewpoint of the phylodynamic studies, the explicit integration of time means that we can make comparable estimates with epidemiological data, such as disease incidence and prevalence (Pybus and Rambaut, 2009).

1.2.2 Trait Evolution

In addition to ascertaining the viral population dynamics and history from sequence data, phylogenetic analyses can be used to investigate the evolution of viral traits that are expected to be epidemiologically important, such as host or location (Lemey et al., 2009, 2010). This approach is part of broader collection called, phylogenetic comparative methods, where the aim is to understand the evolutionary relationships of the traits observed at the tips of the phylogeny (Harvey and Pagel, 1991). Ancestral state reconstruction is often employed to understand trait evolution among fast-evolving RNA viruses. The primary aim is to infer the trait values of the unobserved states in the phylogeny, i.e. at the internal nodes, given the observed trait values of the sampled taxa at the tips. Central to these methods are the model of trait evolution across the phylogeny and the phylogenetic estimation itself (Bollback, 2006; Huelsenbeck et al., 2003; Nielsen, 2001; Lemey et al., 2009). The earlier trait evolution models utilized maximum parsimony reconstruction (Felsenstein, 1985; Pagel, 1999), which by and large have been

superseded by likelihood-based methods that also consider the branch lengths in the phylogenetic estimation (Bollback, 2006; Nielsen, 2001). Moreover, trait evolution can now be explicitly modelled using stochastic processes such as a continuous-time Markov chain or a random walk (Bollback, 2006; Nielsen, 2001; Lemey et al., 2009, 2010). These models have been recently been integrated into the Bayesian phylodynamic framework, giving the added advantage of co-estimating the trait change in units of time, along with the genealogy and demographic history (Lemey et al., 2010).

In viral epidemiology, traits of interest are linked to understanding how they shape the genetic diversity at the population level, which can be a viral phenotype or an ecological trait of the virus. Typical examples of these traits include, drug resistance mutations, antigenic property or classification, host species and geographical location (Zhai et al., 2007; Lemey et al., 2009, 2010). Understanding the geographic correlation of the viral isolates across the phylogeny, i.e. phylogeography, can help us uncover the underlying patterns of transmission of established and novel emerging viruses. These may be maintained by endemic-epidemic cycles, like HCV and dengue, or are zoonotic (animal) viruses that present perpetual risk of zoonotic transmission to humans, like swine influenza A virus and North African rabies virus (Nelson et al., 2011; Talbi et al., 2010). This phylogeographic approach is commonly applied to determine the source population of viral outbreaks and understand the spatial structure of viral diversity (Magiorkinis et al., 2009; Nelson et al., 2011; Rabaa et al., 2010; Talbi et al., 2010; Lemey et al., 2009) .

Since trait evolution is modelled explicitly, we can also quantify the rate of evolutionary change of the concerned trait in the phylogeny (Bollback, 2006; Nielsen, 2001). This is likely to be indicative of the underlying ecological or evolutionary processes that affect the viral trait, such as host movement or natural selection (Lemey et al., 2009). In the case of phylogeography, depending on the sampling scheme of the sequence data, we can empirically measure the rate of geographic spread or migration by evaluating the diffusion process employed in the trait evolution model (Lemey et al., 2010). This highlights one of the major advantages of trait evolution models in phy-

lodynamic studies, demonstrating how additional epidemiological relevant information can be incorporated into evolutionary analyses of viral sequences.

1.2.3 Homologous Recombination in RNA viruses

Homologous recombination describes the process of genetic exchange of corresponding regions of the genome between distinct but related viruses. This can result from the template switching of the RNA polymerase between different genomic strands during viral replication. Reassortment is a related process, which only occurs in RNA viruses with multipartite genomes. Specifically, reassortment describes the exchange of whole segments between viruses, thus producing new genomic combinations. While these two processes are different, they are dependent on the cell being co-infected with different viral strains and have similar problems for phylogenetic investigations. If ignored outright, recombination and reassortment can greatly affect evolutionary estimates such as substitution rates, date of a clade and the phylogenetic structure (Posada and Crandall, 2002; Posada et al., 2002). For the ease of discussion, I will use the term "genetic exchange" when referring to these processes together.

The impact of genetic exchange on the phylogeny is dependent on when the event occurs in the evolutionary history of the sampled taxa (see Figure 1.1) (Wiuf et al., 2001; Posada and Crandall, 2002; Posada et al., 2002). If this occurs before two lineages coalesce, then it will have a discernible result on the phylogeny, where different regions of the genome will yield contrasting topologies (Figure 1.1 C). Conversely, if the genetic exchange event takes place shortly after the lineages have coalesced, it will not affect the overall topological structure of the inferred phylogeny across the genome, however the estimated branch lengths may be affected depending on the interval between the coalescent and non-vertical evolution event (see Figure 1.1 A and B). In phylogenetic studies, this is likely to translate to increase in uncertainty of the evolutionary parameters, although this can be directly estimated with the Bayesian phylogenetic approach. The frequency and the ability to undergo homologous genetic exchange varies widely among and within viral families. This suggests that in RNA viruses these processes

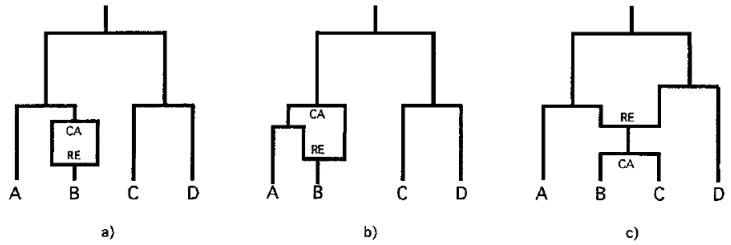


Figure 1.1: A diagram from Posada et al. (2002) (but also refer to the original paper by Wiuf et al. (2001)) to illustrate the impact of recombination event on phylogenetic estimation. Recombination events in a) and b) do not affect the phylogenetic structure, since the lineages have not diverged significantly after the coalescent event. In contrast c) two lineages have been merged by recombination before they coalesced, which will lead to different topology from flanking regions of the breakpoint. In b) the recombination between the different ancestors of A and B, will affect the branch length estimations, while in a) the recombination most likely has occurred between effectively identical sequences, it not affect the branch length estimation.

have not been uniformly selected to decrease the high mutational burden or increase the rate of fixation of advantageous mutations in the population. Different lines of evidences, from both experimental and comparative studies, indicate that the ability to partake in genetic exchange is governed by the constraints imposed by the viral ecology, the host immune response and the genome architecture of the virus (Worobey and Holmes, 1999).

Firstly, the geographic distribution of the virus is an important factor in determining whether hosts are at risk of being co-infected with different viral strains (Worobey and Holmes, 1999). For instance, viruses that are found in isolated regions or populations are unlikely to have the opportunity to recombine or reassort with another related virus. Secondly, the host immune response against the initial infection, depending on whether it is weak or strong, determines if a successful co-infection can occur within the host and the cell (Worobey and Holmes, 1999). Rapid immune clearance or protective immunity can prevent two viruses coming in contact inside the host, regardless of co-infection (Danis et al., 1993). Lastly, once the preceding stages have been established, the genomic structure of RNA viruses seems to be a major determinant of genetic exchange (Worobey and Holmes, 1999; Chare et al., 2003). For example the unusual architecture of retroviruses, as exemplified by HIV, have made them especially amenable to genetic

exchange since it can package two copies of the genome inside the virion. Consequently, in HIV and the ancestral SIV, recombination contributes significantly to maintaining genetic diversity (Lemey et al., 2006; Robertson et al., 1995). Viruses with single-strand positive sense RNA genomes are found to be more prone to recombination than their negative sense counterparts (Worobey and Holmes, 1999; Elena and Sanjuán, 2007). This pattern is explained by the fact that negative sense RNA genomes are associated with nucleoproteins inside the host cell, which acts to stabilise the structure in the volatile environment (Conzelmann, 1998; Worobey and Holmes, 1999; Chare et al., 2003; Elena and Sanjuán, 2007). Since the uncoating of the viral RNA is not necessary for replication, this is thought to restrict the chance of template switching by the RNA polymerase between different strands (Conzelmann, 1998; Worobey and Holmes, 1999; Chare et al., 2003; Elena and Sanjuán, 2007).

In spite of the constraints outlined above, recombinant and reassortant viruses are often observed amongst emerging RNA viruses in the natural environment. This indicates that co-infection with different viral strains is indeed a common part of their epidemiology, which undoubtedly reflects the host population behaviour and demographic changes (Worobey and Holmes, 1999). Moreover, the large population sizes of RNA viruses means that competent and infectious viral progeny will be produced by genetic exchange, even in the face of small likelihoods. It is unclear whether the degree of genetic exchange is selected amongst viruses, or an indirect benefit as a result of their genomic organization, even when it is a central feature of their evolutionary dynamics, e.g. HIV and influenza A viruses (Bonhoeffer et al., 2004; Shapiro et al., 2006b). Further investigation is required in this area of viral evolution, especially to determine whether recombination or reassortment confers any advantages to the virus at the within-host level, such as fixing drug-resistant mutations or facilitating host-adaptation (Pybus and Rambaut, 2009).

1.3 Evolution and Ecology of RNA viruses

Thus far, I have given a broad overview of fast-evolving RNA viruses that have emerged in the human population. However to understand the underlying processes that shape viral emergence and their disease dynamics, I have focused on a select group of human RNA virus pathogens. These are: influenza A viruses, SARS coronaviruses (SARS-CoV), hepatitis C virus (HCV) and dengue virus. The great diversity of RNA viruses, from its genomic architecture to how it is transmitted in the population can be illustrated by these few examples. This array includes viruses that are maintained in multiple animal reservoirs, as well as those that have adapted to circulate exclusively in humans. They also demonstrate the historical range of the viruses, from a recent cross-species transmission, to those with long-association with their human hosts. Therefore, to help understand the motivation and aims of this thesis, I will introduce the evolution and ecology of these chosen viruses below.

1.3.1 Influenza A Virus

Influenza A virus has received much attention over the recent years, which in some part may be explained by the appearance of two novel influenza A virus strains in the human population in the last decade; H5N1 virus from the avian reservoir (Claas et al., 1998; Guan et al., 2002; Xu et al., 1999) and the pandemic H1N1 virus from swine (Fraser et al., 2009; Smith et al., 2009b). These events indicate that influenza A virus is a formidable threat to the global population due to its ability to re-emerge from the animal reservoirs. Influenza B and C are distantly related viral lineages, which also cause respiratory illness in humans. However, in contrast to influenza A viruses, influenza B and C viruses are associated with milder symptoms and are almost exclusive to humans. Influenza A viruses are persistent in several hosts, which include human, swine and avian populations, with the natural viral reservoir being the wild waterfowl (Scholtissek, 1987; Kida et al., 1994; Webster et al., 1992). All three influenza viruses share a multipartite genome structure, such that within each group (i.e. A, B

or C), if a cell is coinfected with different strains, viruses may exchange segments in a process called reassortment. This has been particular relevant evolutionary force in the emergence of influenza A viruses, where the acquirement of new antigenic properties has led to increased access to new susceptible individuals in the population (Webster et al., 1992). The lack of extant animal host reservoirs in influenza B and C viruses means that while reassortment probably does occur, it is unlikely to contribute to its disease dynamics.

The viral genome comprises of eight segments, two that encode the main antigens: haemagglutinin (H) and neuraminidase (N), while the remaining six encode the internal proteins that are important for replication, transcription, host entry and viral assembly. These include three polymerase subunits, two matrix proteins, two non-structural proteins and one nucleoprotein. Influenza A virus is classified according to its antigenic properties, which is represented by H**X**N**Y**, where **X** and **Y** identify the specific subtypes of the two antigens. These range from 1-16 for haemagglutinin and 1-9 for neuraminidase. The full diversity of influenza A viruses has been sampled from the natural hosts, wild waterfowl, while only a restricted set of subtypes (H1N1, H2N2 and H3N2) have been observed in the human population (Webster et al., 1992; Palese, 2004). For example, H5N1 has not successfully emerged in humans and is characterised by limited transmission among individuals, leading to effectively dead-end infection. By contrast, the pandemic H1N1 strain in 2009 rapidly disseminated across the globe, suggesting that there may be some biological constraints on which subtypes can emerge in humans (Palese, 2004). Seroprevalence data before 1918 suggests that the circulating strains were of subtype H3 and H1, supporting a cyclical pattern of subtype emergence, rather than humans being under threat from a full spectrum of subtype combinations from the avian reservoir (Dowdle, 1999; Palese, 2004). The swine-origin H1N1 pandemic has demonstrated that predicting the next pandemic strains, such as which subtype and time of emergence, remains a challenging task. Nevertheless, understanding the ecology of the main hosts, i.e. swine and avian, in relation to humans is paramount to identifying future emerging influenza A viruses (Webster et al., 1992; Palese, 2004;

Smith et al., 2009b).

Epidemic influenza

Influenza A virus circulates as epidemics in the human population, where the strains are derived from the pandemic lineage, and usually continue to infect the population until the next emergence of a novel virus (Webster et al., 1992). The frequency of epidemics differs between the temperate and tropical regions, where viral exports from the latter regularly seed the temperate outbreaks (Nelson et al., 2006, 2007; Rambaut et al., 2008; Russell et al., 2008). Human migration, which has maintained this source-sink dynamic transmission pattern, also contributes significantly to the viral diversity of both within and between seasons via regular importations from tropical or well-connected locales (Nelson et al., 2006, 2007; Rambaut et al., 2008; Russell et al., 2008). The global pattern of genetic diversity in contrast shows surprisingly low antigenic variation. This is mostly attributed to the frequency-dependent selection on immune-escape viral strains, where a rare variant is selected to evade the host immune system (Ferguson et al., 2003; Koelle et al., 2006; Cobey and Koelle, 2008). This may be further accompanied by selective sweeps and genetic bottlenecks, where selection is sufficiently strong to reduce the genetic diversity of all the viral segments (Ferguson et al., 2003; Koelle et al., 2006; Cobey and Koelle, 2008). These processes give rise to a ladder-like phylogenetic structure for the antigenic genes, which represents the continual change of the viral phenotype, i.e. antigenic drift, over time (Fitch et al., 1991, 1997; Bush et al., 1999; Ferguson et al., 2003).

While reassortment has been observed in epidemic influenza, shaping the viral diversity with and between seasons, its exact role in the evolution and epidemiology of the virus is unclear (Nelson et al., 2006, 2007, 2008b,a). There is some evidence that it may facilitate antigenic change similar to pandemic influenza (Nelson et al., 2008b). Furthermore, reassortment has been shown to play a role in generating favourable genomic combinations from ancestral strains with lower fitness (Holmes et al., 2005). However further investigations is necessary, with a deeper sampling of seasonal strains

over longer periods, to fully elucidate the impact of reassortment on epidemic viruses. One of the aims of this thesis is to provide a foundation for future studies examining reassortment amongst influenza A viruses. This includes developing better methodology to understand the patterns of reassortment across the viral genome, ultimately helping to understand whether this evolutionary force is largely stochastic and neutral or is selectively advantageous to the virus.

1.3.2 SARS Coronavirus

SARS Coronavirus (SARS-CoV) is the causative agent of severe acute respiratory syndrome (i.e. SARS), which emerged for the first time in Southern China in late 2002 (Peiris et al., 2003; Rota et al., 2003; Marra et al., 2003). The virus belongs to the Coronaviridae, typically causing respiratory and gastrointestinal infection of mammals and birds. In humans, coronaviruses account for a significant proportion of common cold infections, exceeded only by rhinoviruses (Lai et al., 2007). This indicates that although SARS-CoV represents a novel emergence, cross-species transmission of coronaviruses has occurred on numerous occasions in the human history. Therefore, it is important to understand the ecology of SARS-CoV and related coronaviruses in the wild, to appreciate the future emergence potential of these viruses.

The related SARS-CoV isolates have been sampled from the animal markets and subsequently from bats in Southern China (Kan et al., 2005; Song et al., 2005; Guan et al., 2003; Woo et al., 2006; Lau et al., 2005; Tang et al., 2006; Lau et al., 2010). The genetic diversity and prevalence of the virus in the latter indicates that bats, in particular the Chinese Horseshoe bat species, (*Rhinolophus spp.*) are the natural reservoir for SARS-CoV (Li et al., 2005; Poon et al., 2005; Woo et al., 2006; Tang et al., 2006; Lau et al., 2010). Interestingly, there has been a recent report of SARS-like CoV sampled from a different bat species in Kenya, suggesting a much wider global distribution of the ancestral viral population of SARS-CoV (Tong et al., 2009). Since the bat virus is not able to bind to the human receptors, there are strong doubts whether they are the immediate hosts of human SARS-CoV (Ren et al., 2008; Shi and

Hu, 2008; Graham and Baric, 2010). Indeed, it is very likely that an intermediate animal reservoir has been involved, perhaps comprising of several host species, playing an integral role in spreading the virus to humans.

Following the isolation of almost identical viruses to human SARS-CoV from the animal markets (98% sequence identity), they have been proposed as the amplifying reservoir, which may have contributed to maintaining high levels of the pathogen during the human outbreak (Guan et al., 2003). In particular, the Himalayan palm civets (*Paguma larvata*) have been identified as carriers of SARS-CoV, and with their presumed large population size, they certainly present as credible viral source for the cross-species transmission to humans. However, there is evidence from both epidemiological and molecular studies that this animal reservoir are unlikely to be the direct source of the human outbreak. Firstly, the SARS-CoV positive animals from the markets are recorded as showing visible signs of respiratory illness (Guan et al., 2003). This suggests, like humans, these animal hosts recently acquired the virus. However, the civets could have been the predecessor of the human outbreak if the virus was circulating in the animal markets before 2002. Secondly, the lack of detection of SARS-CoV in animals from the wild or the farm (where they were harboured before being transported to the marked), suggests that the virus is not naturally endemic in this population (Kan et al., 2005). Moreover, the infected animals are likely to have contracted the virus during the transit to or after they arrived in the markets (Kan et al., 2005). Nevertheless, the observation that distantly related carnivores, such as cats and ferrets, are capable of transmitting human SARS-CoV, suggests that the animal markets could have been a significant factor in the establishment of the human epidemic (Martina et al., 2003).

Detailed information surrounding the epidemiology of the early cases is scant. Whether the palm civets and other animal in the markets contracted the virus prior 2002, i.e. before the first reports of the human SARS cases, is unknown. These questions are especially relevant in determining the risk of future emergences of coronaviruses from the animal reservoir. Increasingly, molecular epidemiology studies are discovering new coronavirus species that circulate endemically in the wild, with bat

populations appearing disproportionately in this category (Lau et al., 2010, 2005; Li et al., 2005; Song et al., 2005; Vijaykrishna et al., 2007; Woo et al., 2006). Determining the natural ecology of coronaviruses, in particular between interacting species, may shed light into the origins of SARS-CoV in humans, and help identify the main factors involved in its successful cross-species transmission (Vijaykrishna et al., 2007).

1.3.3 Hepatitis C Virus

Hepatitis C virus (HCV) infection presents a major global health burden, with WHO estimating 170 million chronic carriers at risk of developing severe clinical liver diseases such as cirrhosis and hepatic cellular carcinoma. The virus belongs to the RNA virus family *Flaviviridae* and is characterised by considerable genetic diversity. HCV is classified into six main genotypes (1-6), each of which are further divided into numerous subtypes. The virus exhibits nucleotide sequence divergence of 30 and 20% at the genotype and subtype levels, respectively. The high level of genomic heterogeneity of HCV is a result of both its high rate of evolution and its long-term association with human populations (e.g. Smith et al. (1997)). Although a zoonotic source remains to be determined, a related virus has recently been discovered in dogs (Kapoor et al., 2011).

The current distribution of HCV genotypes and subtypes is geographically structured, reflecting differences in the rates and routes of transmission of the various subtypes and genotypes. The greatest diversity of HCV viruses is found in West and Central Africa and in Southeast Asia, where the virus appears to have persisted endemically for at least several centuries (Pybus et al., 2007a; Smith et al., 1997). Epidemic strains, exemplified by subtypes 1a, 1b, and 3a, are characterised by high prevalence, low genetic diversity, and a global distribution. These viruses are typically associated with transmission via infected blood products and injecting drug use (IDU) during the twentieth century (Dubois et al., 1997; Pawlotsky et al., 1995; Pol et al., 1995; Pybus et al., 2001; Schreiber et al., 1996; Seeff et al., 2000). In contrast, endemic strains are more spatially restricted but display greater genetic diversity than epidemic

strains. Moreover, the epidemic strains, which constitute to the majority of HCV infections worldwide, are thought to be imported from areas where HCV is endemic (Pybus et al., 2005; Smith et al., 1997).

HCV is somewhat an unconventional pathogen for fast-evolving RNA viruses typically associated with acute infectious diseases. With the exception of retroviruses, HCV and the distantly related GBV-C flavivirus can lead to chronic infections. In addition, HCV may represent a much older human disease, originating before humans migrated out of Africa. Moreover, it is likely to have persisted in small hunter-gatherer population due to the long infection periods. Dating analyses have inferred times of divergence between 500-2000 years for all genotypes (Pybus et al., 2001; Smith et al., 1997). Although, historical reports of hepatitis and the widespread distribution of HCV in different populations around the world, including even amongst isolated groups such as pygmies in Cameroon, suggests that this timescale is likely to be conservative (Kurbanov et al., 2005). A recent study that investigated the age of an endemic HCV strain in East Asia estimated a larger time interval, 600-3000 years, supporting a much older divergence for HCV diversity (Pybus et al., 2009).

Both host persistence and the underestimation of the long-term evolutionary rates appears to be linked to the genome-scale ordered RNA structure (GORS)(Simmonds et al., 2004; Tuplin et al., 2002). This is the observation of extensive internal base pairing in the viral genome. Although the exact mechanism is uncertain, it has been proposed to help evade the host defence system by preventing immune recognition (Simmonds, 2004; Simmonds et al., 2004). Similar patterns are observed in the vesivirus and the foot and mouth disease virus (Simmonds et al., 2004). Interestingly, GBV-C also shares these genomic and clinical traits (Tuplin et al., 2002), indicating that GORS may have arisen early in Flaviviridae history, at least since the last common ancestor of HCV and GBV-C.

The presence of GORS in HCV also means that the number of neutral sites is drastically reduced in the genome, such that homoplasy and convergent evolution are expected to be common in its molecular evolution (Simmonds, 2004; Simmonds et al.,

2004; Tuplin et al., 2002). This suggests that standard model of nucleotide substitution are likely to yield underestimates when inferring long-term divergences in HCV due to their impracticality to deal with high frequency of multiple hits (Simmonds and Smith, 1999; Simmonds, 2004; Smith et al., 1997).

Recombination, like with many other flaviviruses, has been frequently reported in HCV both within and between genotypes. Although the rate of recombination is not high as observed in HIV, the increasing number of studies detecting natural HCV recombinants in the population suggests it is not entirely insignificant force in the viral evolution. Since it has been observed in a broad range of related viruses, including human pathogens such as dengue and encephalitis viruses, along with pestiviruses which mostly affect animals, there does not seem to be any mechanistic constraint to impede genetic exchange in this virus family. The varying rates and observation of recombination amongst members of the flaviviruses most likely reflects differences in the ecological opportunity of these viruses, which will be determined by both transmission modes and host demography. It also could be explained by the under-sampling of the true viral genetic diversity, where recombinants are expected to occur low frequency in the population.

The slow discovery of recombinants indicates that recombination is an uncommon event in HCV, particularly compared to HIV. The most notable HCV recombinant to date is the 2k/1b strain, discovered first in 1999 in St Petersburg, as the only circulating recombinant form (CRF) of HCV. A CRF is a recombinant that is found repeatedly in different patients, where the 2k/1b strain is formally defined as CRF01_1b2k. Subsequent isolations of CRF01_1b2k from different locations in Eurasia, from Ireland to Russia, demonstrates its wide circulation in the population, raising questions about the epidemiological background in which it first appeared. The genesis and dissemination of CRF01_1b2k is investigated in Chapter 4 to understand why it might be unique, and evaluate the likelihood other whether other HCV recombinant forms that could increase in prevalence in the future.

1.3.4 Dengue virus

Dengue is the only arboviral disease to have established endemic transmission in humans. It is found in the subtropical and tropical parts of the world and is caused by a group of four related viruses, which all belong to the single-stranded positive sense family*Flaviviridae* (Initative, 2009; Gubler, 1998). Each dengue virus represents an independent introduction into the human population, which are denoted by their serotypes DENV1 to DENV4. The disease is transmitted by the *Aedes* mosquitoes, where the two main vectors in humans are *A. aegypti and A. albopictus*(Gubler, 1998). In particular, *A. aegypti* has successfully adapted to the human host and the urban environment, where dengue is most prevalent (Gubler, 1998).

The natural viral reservoir for dengue is found amongst the Old World Monkeys in the forests of Southeast Asia and Africa, where a separate sylvatic transmission cycle exists via a different *Aedes* mosquito species (Gubler, 1998). Spillover infections from this sylvatic source (i.e. dengue viruses that typically circulate exclusively in the wild animal populations) to humans are occasionally reported, providing a plausible route for future emergences of new dengue viruses in the population (Vasilakis et al., 2011). The origins of the human dengue viruses is surprisingly recent, of the order of few hundred years for each serotype (e.g. DENV1-4), with all four strains diverging approximately 1000 years ago (Holmes and Twiddy, 2003; Twiddy et al., 2003). The sustained transmission of dengue, such that the virus can circulate endemically in the human population without the requirement of a zoonotic source, has only been established around 100 years ago (Holmes and Twiddy, 2003; Twiddy et al., 2003). This period corresponds to major transitions in human ecology, such as exponential population growth of societies, increasing urbanisation and human movement, which are likely to have favoured the endemic transmission of dengue (Gubler and Meltzer, 1999).

The rising global prevalence of dengue, especially over the last 50 years, combined with increased severity and frequency of dengue outbreaks, has become a major concern for public health (Gubler, 1998; Gubler and Meltzer, 1999; Guzmán and Kourí,

2002). The exact cause for these shifts in dengue epidemiology are unknown, although demographic and behavioural changes of both vector and human populations are central to the processes shaping the viral transmission (Gubler, 1998; Guzmán and Kourí, 2002). The population and geographic expansion of the mosquito vectors can explain the increasing dengue incidence and morbidity. In combination with changes in human activity and behaviour, such as large-scale movement and growing urban areas, *Aedes* mosquitoes have been able to be maintained in large population sizes and introduced to new parts of the world (Gubler, 1987, 1998; Gubler and Trent, 1993; Guzmán and Kourí, 2002). For example, the human movement during the last world war has been identified as one of the main reasons for shaping the current geographic distribution of the vector population. Moreover, it is clear that urbanization has greatly influenced endemic transmission of dengue in humans, by firstly providing ample supply of breeding sites (i.e. stagnant water) and ultimately a high density of susceptible hosts (Gubler, 1987, 1998; Gubler and Trent, 1993). Although, initial dengue infection confers lifelong immunity to that viral serotype, subsequent infections are common where different serotypes circulate since little cross-protective immunity exists against all four viruses (Gubler, 1998). The re-exposure to different dengue strains is associated with more severe clinical outcomes, namely dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS) (Guzmán and Kourí, 2002). These conditions can be fatal and occur disproportionately among young children (Guzmán and Kourí, 2002). Host and viral factors are also known to contribute to developing these critical conditions (Guzmán et al., 2000; Guzmán and Kourí, 2002). The prevalence of DHF/DSS is greatest where multiple strains are endemic, and is further associated with the introduction of new dengue serotype into previously exposed populations (Gubler and Trent, 1993; Gubler, 1987). Since most dengue infections are asymptomatic, it seems inevitable that individuals will encounter different dengue viruses, raising the likelihood of DHF/DSS in the population (Gubler and Meltzer, 1999; Gubler, 1998; Gubler and Trent, 1993; Gubler, 1987; Guzmán and Kourí, 2002).

The lack of vaccine to protect against all four serotypes, combined with dengue

predominantly affecting the developed countries, has created a great health and economic burden in places where it is endemic (i.e. areas with regular outbreaks). The main strategy to manage the disease has been to target the vector population directly, either by pesticides or environmental planning. However, for effective deployment of vector control strategies and intervention plans, a better understanding is required in how dengue transmits in endemic populations, including identifying the central factors involved. To this end, the viral transmission in Southeast Asia has been investigated at different geographical scales in Chapter 6 to evaluate the spatial and temporal dynamics of dengue in a region with the highest global prevalence.

Thesis Outline

In the next chapters, I address some of the challenges presented by fast-evolving RNA viruses when investigating their evolutionary history and molecular epidemiology. In particular I have concentrated on the following areas: 1) improving phylogenetic methodology to incorporate recombination and reassortment into phylodynamic studies, 2) investigating the evolutionary origins of emerging RNA viruses in face of epidemiological or evolutionary constraints, such as a novel cross-species transmission and ancient viral emergences, and 3) spatiotemporal dynamics of emerging viral diseases.

Chapter 2 introduces a new phylogenetic-based tool to examine reassortment in influenza A viruses. While this evolutionary force is understood in the context of pandemic influenza, its impact on seasonal influenza is unclear. Here, I describe an efficient algorithm to find parts of the phylogenies, constructed from different gene segments, that are in common. Therefore, by understanding the degree to which these evolutionary histories differ across the genome, we can obtain a measure of reassortment.

Chapter 3 questions the origins of the SARS outbreak from the currently sampled animal reservoirs, palm civets and bats, as proposed by previous studies. To address this controversial topic, I evaluate the role of palm civets and bats in a statistical robust Bayesian phylogenetic framework. I demonstrate that while we have isolated related SARS-CoV from these hosts, there is not strong evidence to implicate them as the

predecessor of the human SARS-CoV. The main reasons for the lack of uncertainty appears to in line with restricted temporal viral samples from the animal markets and the under-sampled diversity of the bat SARS-CoV.

Chapter 4 investigates the genesis and transmission of the only known circulating recombinant form of HCV, CRF01_1b2k. This work has been carried out in collaboration with authors from Amsterdam, who provided new CRF01_1b2k sequence samples, and Japan. A hierarchical Bayesian approach was employed to jointly estimate the origin of CRF01_1b2k from different genomic regions. Consequently, the findings provide a credible role of the former Soviet Union and the worlds first centralized national blood service in the emergence and geographic spread of the unique HCV CRF.

Chapter 5 examines the long-term association of influenza viruses (A, B and C) with humans. While current evolutionary models have been used to investigate the short-term evolutionary dynamics of emerging RNA viruses, it is not clear if they are appropriate for investigating long evolutionary timescales, that may span over thousands or millions of years. Here I employ an analytical framework, utilising both nucleotide and amino-acid substitutional models, to date the divergence of the distantly related influenza viruses.

Chapter 6 looks at the spatiotemporal patterns of endemic dengue transmission in Southeast Asia over different geographic scales. This study exploits a large viral sequence data set sampled from southern Viet Nam, which was collected and sequenced by Cameron Simmons and colleagues in Ho Chi Minh City and the Broad Institute. By performing a comprehensive analysis on this data set, as well as related viruses from neighbouring regions, using Bayesian phylogeography models, we demonstrate the complex interplay between spatial, genetic and epidemiological dynamics of DENV-1 at local, regional and global levels. The findings suggest that the host population density is key driver of viral transmission at local and regional scales. Furthermore, at the local level, we find surprisingly low viral movement in both urban and rural environments, indicating that dengue transmission is largely mediated by mosquito dispersal, rather than human movement.

2

Quantifying Reassortment in segmented viruses

2.1 Introduction

Reassortment in influenza A viruses can be a powerful evolutionary force, particularly in generating pandemic-potential strains by swiftly changing viral antigenic properties. However, the impact of reassortment on seasonal influenza diversity is less well understood, although recent reports indicate that it may have a similar role in producing novel antigenic combinations (Nelson et al., 2006, 2008b,a). Furthermore, the process of segmental exchange may facilitate the emergence of epidemic-potential lineages by allowing viruses to explore favourable genomic combinations, thus increasing their fitness relative to contemporaneous strains (Holmes et al., 2005). The recent emergence of the human pandemic strain, swine-origin H1N1, illustrates that reassortment can greatly obscure the evolutionary history of influenza A viruses (Smith et al., 2009b). In spite of all gene segments originating from a swine source, inspection of the evolutionary history revealed that the H1N1 lineage was a product of multiple reassortment events involving: swine, human and avian viruses (Smith et al., 2009b). A recent study that examined the long-term evolutionary and transmission dynamics of swine influenza A viruses in Hong Kong has found substantial amount of reassortment in the swine population (Vijaykrishna et al., 2011). This observation of great viral diversity in the swine reservoir suggests that there were many possibilities for the emergence of the pandemic strain in 2009. Therefore, due to the ecological interaction between humans and swine, reassortment is likely to contribute to generating future viral strains that could affect human populations. However, our current knowledge and methods to understand reassortment are limited. A common approach to is to identify incongruent lineages between phylogenies estimated from different gene segments (Nelson et al., 2008a,b, 2006; Nagarajan and Kingsford, 2011). Therefore, fundamental questions about the pattern and extent of reassortment across the viral genome and in different populations have yet to be addressed.

To understand reassortment in epidemic influenza, I introduce a new methodology to help quantify this process among a set of sampled taxa, where multiple gene seg-

ments have been sequenced. Given a pair of phylogenies that have been constructed from different parts of the viral genome, the frequency of reassortment will determine the degree to which the topological structures will vary between gene segments. Consequently, this relationship can be exploited to find out about the extent of evolutionary association, and thus reassortment, across the viral genome. To measure reassortment, the strategy I have adopted first relies on finding the parts of the trees that are common to both. This approach is similar to a well-known problem in graph theory called the maximum agreement subtree (MAST) (Finden and Gordon, 1985). For a given set of trees, the MAST is the largest common subtree (i.e. with the same topological structure) that is found in each tree, containing the same taxa. Heuristic algorithms to find the MAST have been proposed for evolutionary trees that are rooted and bifurcating. These particular trees are classified as graphs with a bounded degree, indicating that the internal nodes have a restricted number of directed edges (i.e. branches). For graphs with unbounded degrees, obtaining the MAST is a NP-complete problem, where finding an exact solution in realistic time is thought to be impossible (Amir and Keselman, 1997). However, for the constrained rooted binary trees, there are efficient search methods to find the MAST or MASTs for a given set of trees (Cole et al., 2000; Lee et al., 2005). Nevertheless, the performance of these algorithms ultimately depend on the number of taxa included in the tree set, since this determines the search space of the number common subtrees. Thus, for trees with a large number of taxa, i.e. when n ¿ 100, these algorithms are impractical as the time to find the MAST increases combinatorially. The phylogenetic uncertainty is also expected to increase with large taxa sets. This chapter specifically addresses the difficulty of comparing large phylogenies and extracting the common evolutionary history across the genome. A recent method using MASTs to summarize the posterior tree distribution has been proposed by Cranston and Rannala (2007). However, this approach is not applicable to comparing tree distributions estimated from different genomic regions, which is the central aim of the algorithm presented in this chapter.

An efficient algorithm to ascertain the largest common subtree has been developed

in Java, which I will describe in more detail in the following sections. By applying this algorithm to a pair of tree distributions that have been estimated from different genomic regions, the degree to which reassortment has affected the evolutionary histories can be determined with statistical support, either in the form of bootstrap values or posterior probabilities. In contrast to previous approaches that have looked at quantifying reassortment in influenza A virus genomes, which only identify incongruent taxa between a pair of phylogenies (Nagarajan and Kingsford, 2011; Yurovsky and Moret, 2011), the method introduced here has additional advantages by extracting the long-term evolutionary history or the common evolutionary backbone that may be shared between genes. Two such examples where common subtrees can help gain deeper insight into the evolution of influenza A viruses are: 1) antigenic evolution: determining to what extent reassortment or adaptive evolution are important, and 2) epistatic interactions between genes; where evolutionary change occurs concertedly among independent segments due to some functional interdependence at the protein level. Therefore, besides serving as a tool to investigate and quantify reassortment, the common subtree application can help examine the detailed evolutionary patterns that may be affected by segmental exchange.

2.2 Methods

Brief overview

Since reassortment will affect the topological structure of the phylogenetic trees constructed from different genomic regions, the Robinson-Foulds metric (Robinson and Foulds, 1981) was employed to determine the degree of similarity between a pair of trees. This metric decomposes the trees into the constituent subtrees and identifies which are unique to each tree. The sum gives us a measure of how many incongruent clades there are among the pair of trees. As this metric does not take into the hierarchical structure of the phylogeny, taxa can appear multiple times in the incongruent clades so the topological difference can be overestimated. To address this issue, an

algorithm was developed to determine the largest common subtree between a pair of rooted, bifurcating and labelled trees.

The Common Subtree algorithm

For trees based on a large number of taxa, the search space to find the common subtree can be combinatorially large, such that exhaustive solutions are computationally impractical. Below, I describe the search strategy adopted in this chapter, which uses a greedy algorithm to find the minimal number of taxa to remove to obtain a common subtree. The main idea behind the algorithm is to compare the clades of the nodes in the trees in the order of tree depth (distance from node to tip) from root to tip, with the aim of retaining the greatest number of taxa in the clades. To help with this strategy while traversing the tree, six different comparisons are calculated at each node (see Figure 2.1 B), where the action resulting in the fewest number of taxa being pruned is chosen. If we refer to the schematic of the algorithm in Figure 2.1, starting from the root node, we consider the descendants (the child nodes) along the left and right subtrees, i.e. L and R. Since the child nodes can rotate with respect to the parent node, the assignment of left and right subtrees is arbitrary. However, by performing six comparisons, the algorithm does not need to search through the different rotations of the node, as an explicit labelling of the child nodes is employed. As the algorithm traverses the tree, it calculates scores according to the six comparisons illustrated in Figure 2.2. The score is based on the sum of taxa or tips that are present in the common clades between the trees. If we take the example presented in Figure 2.1, comparing the two left subtrees (LL) and the two right subtrees (RR) from the tree pair, gives the highest number of taxa in the common clades (4 and 3 respectively) (Figure 2.2). Therefore, the best score for the first move, LLRR, is 7 (see Figure 2.2).

In the case of equal best scores, the ties are broken by randomly selecting a move. This strategy of finding the best score continues until the algorithm reaches the tips, and the two trees have been pruned to have identical topological structures. Since the number of pruned taxa to obtain the common subtree depends on the tree traversal

Figure 2.1: A schematic to explain the scoring system used to compare two rooted bifurcating trees with identical taxa set. The algorithm is recursive, which starts at the root node, comparing six different clades between two trees for each internal node, and stops once it reaches the tips. A) If we begin at the root node (yellow), there three clades that we can compare for each tree. These are the left and right subtrees of descendant nodes (denoted as L and R in the diagram), as well comparing the entire tree or clade (i.e. N). Given this, B) there are six possible comparisons to be made between Tree 1 (red) and Tree 2 (blue). The number of taxa that are shared between each comparison gives the score for each move. For example, the score for LLRR (i.e. comparing the two left subtrees, LL, from Tree 1 and Tree 2, as well as comparing the two right subtrees, RR, from Tree 1 and Tree 2), will be LL = 4, since A, B, C, and D are common in the left subtrees, while RR = 3, since F, G and H are common in the right subtrees. Therefore the total score for LLRR = 7. This is calculated for each move, where the best score is the one that retains the highest number of taxa.

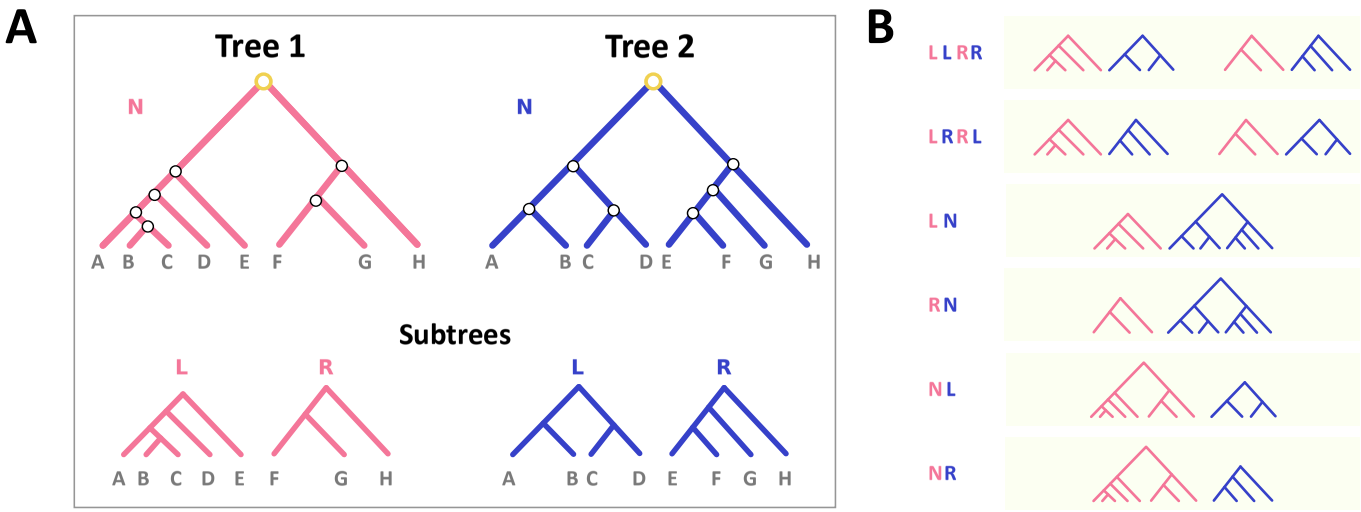


Figure 2.2: An example to illustrate how the common subtree algorithm works on the rooted bifurcating tree pair introduced in Figure 2.1. Once the initial move has been calculated from comparing the clades from the root node, either the whole clade (N), the left subtree (L), or the right subtree (R), the recursive algorithm moves down the node that retains the highest number of taxa between Tree 1 (red) and Tree 2 (blue). Since the Since the initial move is LLRR, the algorithm will traverse down the two left subtrees. This traversal (and pruning) of the two left subtrees continues until the tips (left box). Once the two left subtrees are pruned to be topologically the same, the algorithm moves on to comparing the two right subtrees (right box). Importantly, any taxa pruned during the traversal, they are subsequently pruned from all trees. Hence, the two right subtrees are now both ((F,G) H) since E has been already pruned in the left box.

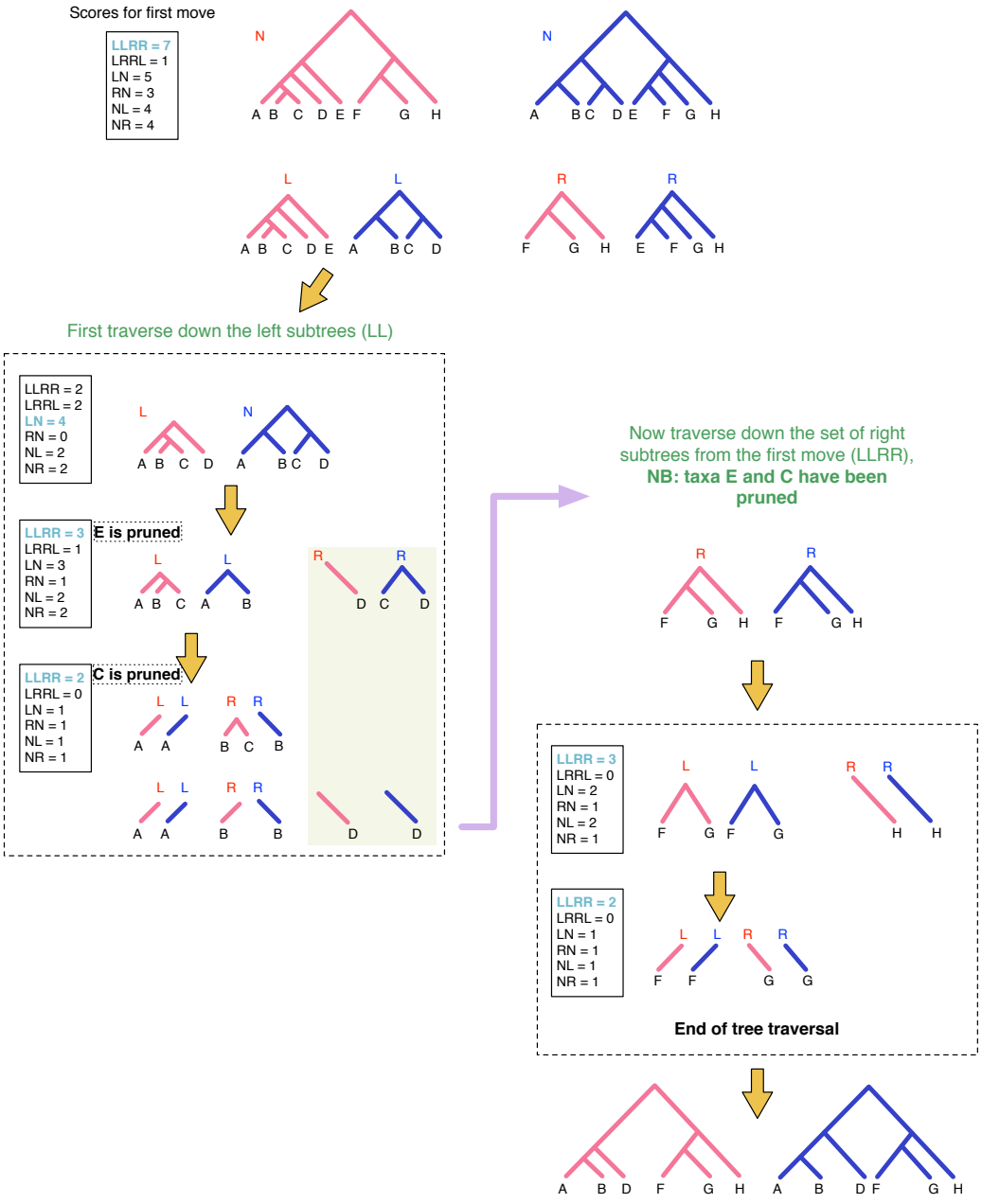


Table 2.1: The execution time results for the algorithm when tree distributions are compared for increasing number of taxa

Number of taxa (n)	Execution time (millisecs)
21	38795
133	214805
687	991363

pathway, which can differ due the nodes with tied best scores, the algorithm executes 1000 random restarts to find an optimal solution. Figure 2.2 illustrates the tree traversal and pruning strategy for a pair of hypothetical trees introduced in Figure 2.1. The important points to note are: 1) the traversal algorithm continues down the chosen subtrees until it has reached the tips, and 2) any tips that are removed during the traversal down one route, are consequently removed from all clades in the pair of trees. For example, in Figure 2.2 the first move is LLRR, which means the set of left subtrees will be compared first before the set of right subtrees. During the traversal of the left subtrees, two tips are removed, one of which (E) is present in the set of right subtrees, and will be pruned before the algorithm returns to compare the set of right subtrees.

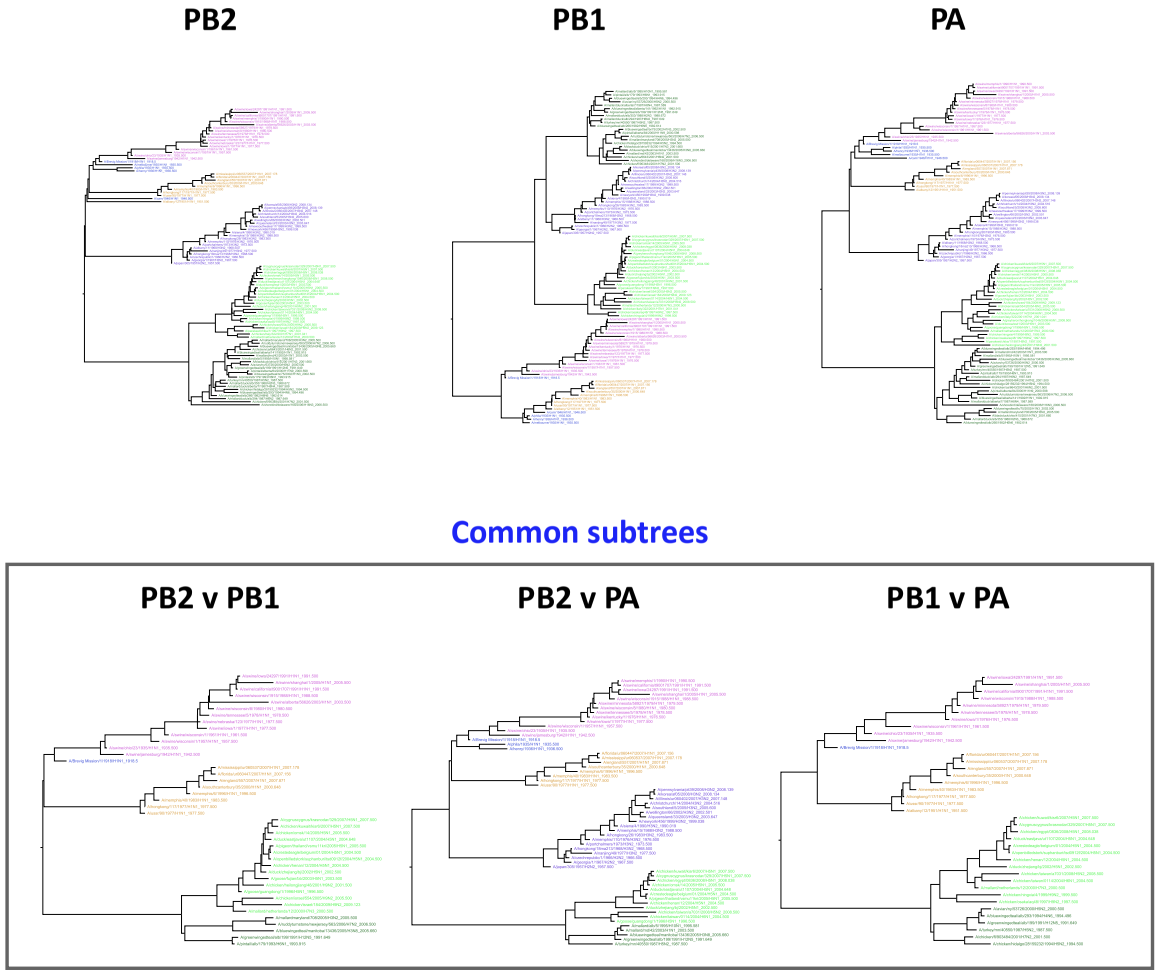
2.3 Results and Future Work

The algorithm was found to work efficiently when comparing a pair of trees with large number of taxa, e.g. for n = 291 the common subtree was computed in less than 3 seconds. In the case of analysing tree distributions, where pair of trees are compared sequentially from a posterior sample based on 10,000 estimates, the execution time of the algorithm increased linearly with the number of taxa (Table 2.1). This relationship most likely reflects the number of restarts is fixed in the algorithm, irrespective of the taxa size, which could be improved to scale with increasing number of taxa.

To illustrate the performance of the algorithm on empirical data, I tested it against some results obtained in the subsequent chapters. The example in Figure 2.3 is based on the polymerase trees of influenza A viruses sampled from avian, swine and humans from Chapter 5. The trees presented in Figure 2.3 include 97 isolates to represent the

influenza A viral diversity, which were estimated from a posterior distribution using BEAST (Drummond and Rambaut, 2007) (see Methods in Chapter 5 for more detail). A pairwise comparison between the polymerase trees found the human H3N2 viruses (colour labelled purple in Figure 2.3) to be reassortant viruses, where PB1 came from the avian reservoir, while the other two polymerase genes was derived from the previous circulating human seasonal flu virus. Since the chapters 3 and 4 look at recombination in SARS-CoV and HCV respectively, the posterior tree distributions estimated from flanking regions of the breakpoint provide good test data for the algorithm. In the case of SARS-CoV, two recombinant bat SARS-CoV lineages identified by Hon et al. (2008) and Yip et al. (2009) were correctly removed when the algorithm was applied to the trees presented in Figure 3.2 (Chapter 3). When I compared the clade representing the HCV circulating recombinant form (CRF) 2k/1b, the algorithm removed most of taxa to obtain a common subtree (on average 21/27, see Figure 2.4 D). To ascertain how many tips would be removed by chance, the tips of the tree distributions were randomized before applying the algorithm (Figure 2.4 C). Since this removed a similar number of taxa, the actual results are likely to reflect the phylogenetic uncertainty associated with the short sequences, rather than multiple recombination events. Even when comparing independent tree distributions, estimated from the same genomic regions, a large proportion of the taxa were removed, which would be expected if phylogenetic uncertainty is sufficiently high (Figure 2.4). While the groundwork has been carried for this chapter, further testing is required to see how it compares with available methods, such as GiRaF: graphical incompatibility based reassortment finder (Nagarajan and Kingsford, 2011), and whether additional improvements could be made to the algorithm. The fact that this method has been directed towards comparing time-scaled phylogenies does mean it should be more powerful in measuring reassortment in influenza A viruses, since as with rate of evolution, these processes are a product of time. Future work also needs to address the best way to summarise the results when comparing a pair of posterior tree distributions. Although tree distributions can be efficiently stored as a nexus file when it is based on the same taxa set, describing the common subtree

Figure 2.3: Results when comparing the polymerase trees of influenza A viruses from avian, swine and humans. The algorithm correctly determines the Human H3N2 lineage (purple) as reassortant in the PB1 tree. Colour labels for the clades: Classical Swine - pink; Human H1N1 (1918-1957) - blue; re-emergent 1977 Human H1N1 - orange; Human H3N2 - purple; Eurasian Avian - light green; North American Avian - dark green.



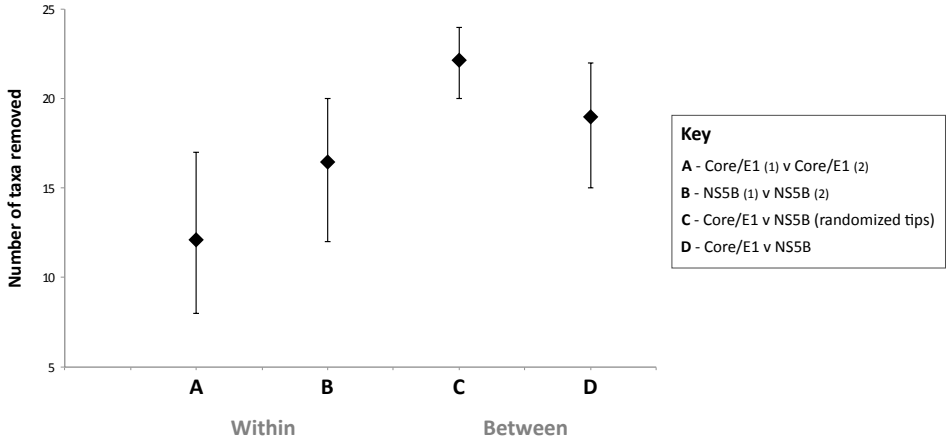


Figure 2.4: Results from comparing tree distributions estimated from different genomic regions (Core/E1 and NS5B) of HCV CRF01_1b2k in Chapter 4. These results represent comparisons of tree distributions estimated independently from the same genomic regions ("Within" - A and B), as well as comparing tree distributions estimated from different genomic regions ("Between" - C and D). In comparison C, the tips of the tree distributions were randomized in order to calculate how many taxa would be removed by chance and if there is no phylogenetic uncertainty.

distributions has the added complexity since the number of tips and the included taxa set will vary. Lastly, while the focus here has been on investigating a particular type of homologous genetic exchange in RNA viruses, the method could be applied to examine phylogenetic estimations along the genome where recombination is prevalent, such as HIV.

3

Origins and Cross-species Transmission of the SARS outbreak

3.1 Introduction

The emergence of SARS coronavirus (SARS-CoV) demonstrates the threat that humans face from fast-evolving viruses that jump the host species barrier. In addition, the rapid global dissemination of SARS, similarly observed with the recent swine-origin H1N1 pandemic (H1N1pdm), highlights the importance of human movement in shaping the viral epidemiology (Chinese SARS Molecular Epidemiology Consortium, 2004). The SARS epidemic, in particular, is characterised by a series of super-spreader events (SSE), where one or few infectious individuals transmit the disease agent to a large number of individuals, which have been associated with hospitals (Chinese SARS Molecular Epidemiology Consortium, 2004). The combination of novel origin and transmission patterns of SARS-CoV during 2002-2004, have provided an interesting case for both evolutionary biologists and epidemiologists to investigate and gain further understanding of emerging viral diseases.

The role of a zoonotic source has been considered since the early reports of SARS in late 2002 (Zhong et al., 2003). Accounts of animal handling in the first SARS cases and subsequent isolation of closely related viruses from the animal markets have put the palm civets in the spotlight in the sudden appearance of SARS in the human population (Guan et al., 2003; Zhong et al., 2003). For reasons discussed in Chapter 1, it is still uncertain whether this host species has played a significant part in transmitting SARS-CoV to humans (Kan et al., 2005). The later discovery of the SARS-like CoV in bats has also raised the question whether this animal reservoir are the direct ancestors of human SARS-CoV (Tang et al., 2006; Woo et al., 2006; Yuan et al., 2010; Poon et al., 2005; Ren et al., 2006; Hon et al., 2008; Lau et al., 2010; Tong et al., 2009; Li et al., 2005). The great genetic diversity observed amongst the currently sampled bat SARS-like CoV strains and the apparent asymptomatic infection indicates that bats are the original hosts of SARS-CoV. Furthermore, a similar pattern has been observed with other related coronaviruses, where bats harbour the greatest viral diversity, strongly suggesting that bats are the natural reservoir for all coronaviruses (Cui et al., 2007;

Vijaykrishna et al., 2007; Gouilh et al., 2011; Tong et al., 2009). Interestingly, the recent identification of a recombinant bat SARS-like CoV has brought the estimate of the human and bat virus divergence to only a few years before the epidemic at around 1998, presenting the bat reservoir as a potential immediate source of the SARS outbreak (Yip et al., 2009; Hon et al., 2008; Lau et al., 2010). Since both civets and bats have been consumed in southern China, both for culinary delicacy or medical purposes, either of these hosts could have credibly transmitted the virus to humans (Chinese SARS Molecular Epidemiology Consortium, 2004; Zhong et al., 2003). Moreover, the exploitation of the phylogenetically conserved receptor, angiontensin-converting II enzyme (ACE2) (Li et al., 2003), may have facilitated the emergence of SARS-CoV in humans from distantly related species, with either little or no viral adaptation. For example, evidence that the human SARS-CoV is able to infect cats and ferrets and maintain secondary transmission indicates that the virus can infect a broad range of potential hosts without the need to substantially change its antigenic surface proteins (Martina et al., 2003). Thus, to investigate the origins of the SARS outbreak, whether civets or bats are the immediate ancestors of human SARS-CoV, a Bayesian phylogenetic approach is undertaken to analyze temporally sampled whole genome sequence from the three hosts (i.e. humans, bats and civets). Different hypotheses for the evolutionary history of human SARS-CoV are tested in a statistically robust framework, allowing the appropriate evaluation of animal reservoirs in the recent and future emergence of coronaviruses.

3.2 Methods

SARS-CoV Data

An initial dataset of temporally sampled whole genome SARS-CoV sequences (30,000nt) was collated from Genbank, retaining viral isolates where temporal information was readily available. Further dates were procured either from the literature or via personal communication (for the isolates included in this study, the full information is available in Appendix A - Table A.1). A multiple alignment for the whole genome

was constructed manually, where regions with ambiguous homology, indicative of indels or saturation, were removed on the basis of their low information content. The final alignment included the major ORFS (ORF1ab - RNA polymerase, spike - the envelope protein, matrix protein and nucleoprotein) and accessory proteins (ORF 7, ORF 8 and ORF 9). Since the human viral sequences had been oversampled with respect to the other two hosts, a preliminary neighbouring tree analysis was carried out in PAUP (version 4) (Swofford, 2003), using a HKY85 substitutional model, on the original dataset. This helped to identify isolates from epidemiologically-linked clusters that originated from a single point of infection. This is particularly relevant for the human SARS outbreak where SSEs played a major part in its global spread. These were subsequently downsampled, whilst conserving the phylogenetic structure, using a random selection method. The final set comprised 76 isolates: 18 bat, 15 civet and 54 human sequences.

Recombination

The dataset was analysed with GARD (Genetic Algorithms for Recombination Detection) and SBP (Single Breakpoint analyses) programs (Kosakovsky Pond et al., 2006), available from www.datamonkey.org, to investigate previously identified breakpoints by Hon et al. (2008).

Phylogenetic Analyses

Epidemic SARS-CoV

To assess the role of civets in the emergence of human SARS-CoV, a subset of the data, excluding the bat SARS-CoV sequences, was analyzed in BEAST by employing a relaxed uncorrelated lognormally-distributed clock (UCLN) model (Drummond et al., 2006), a codon structured SDR06 substitutional model (Shapiro et al., 2006a) and a GMRF (Gaussian Markov Random Field) coalescent model, which employs a time-aware smoothing prior to infer the population size dynamics (Minin et al., 2008). Since the human viruses clustered into two monophyletic groups, corresponding to the differ-

ent phases of the epidemic, early or late, these sequences were considered accordingly. For isolates where a specific sampling date was not available, i.e. in day, month and year, a uniform prior was applied for the time range with informed bounds according to the literature. The MCMC (Markov Chain Monte Carlo) chains were executed for 200 million generations and sampled at every 20,000$^{\text{th}}$. Multiple runs were performed to evaluate that the chains converged on the same sampling distributions. A prior tree distribution was also estimated for this dataset by re-running the BEAST analysis outlined previously without the sequence data. The evolutionary relationship of the epidemic strains, in particular the human viruses, was evaluated by performing a bayes factor (BF) test on the prior and posterior probabilities for the three mutually exclusive hypotheses (see Figure 3.1). Specifically, the BF values were calculated as the ratio of the posterior and prior odds for each model comparison.

Human and Bat Divergence

The findings from three recent studies (Hon et al., 2008; Lau et al., 2010; Yip et al., 2009), which dated the human-bat SARS-CoV lineage to four years before the epidemic in 2002, motivated a rigorous Bayesian phylogenetic analysis on a subset of the data. This included the currently sampled bat SARS-CoV sequences, along with 2 human and 2 civet sequences to represent the epidemic group. Since a recombination breakpoint was supported by both GARD and SBP analysis, at the ORF1ab and spike gene boundary (between 21494-21495nt), as established by Hon et al. (2008), the sequence dataset was partitioned around this breakpoint and examined separately with BEAST (Drummond and Rambaut, 2007). To estimate the age of the human-bat lineage and investigate further the model choices of previous studies, both constant population and Bayesian skyline prior were employed under a range of relaxed clock models, i.e. uncorrelated gamma (UCGD), lognormal (UCLN) and exponentially distributed (UCED) (Drummond et al., 2006). The last clock model (UCED) is expected to be biologically unrealistic for modelling rate variation among branches in the phylogeny, since it presumes that most branches are evolving very slowly, with some evolving exceptionally

fast. However, it was investigate in this study to evaluate the prior choices employed by Lau et al. (2010); Hon et al. (2008); Yip et al. (2009), which estimated a recent date of divergence for the human and bat SARS-CoV lineage, placing the bats in the probable range as the immediate ancestor of the human outbreak.

3.3 Results

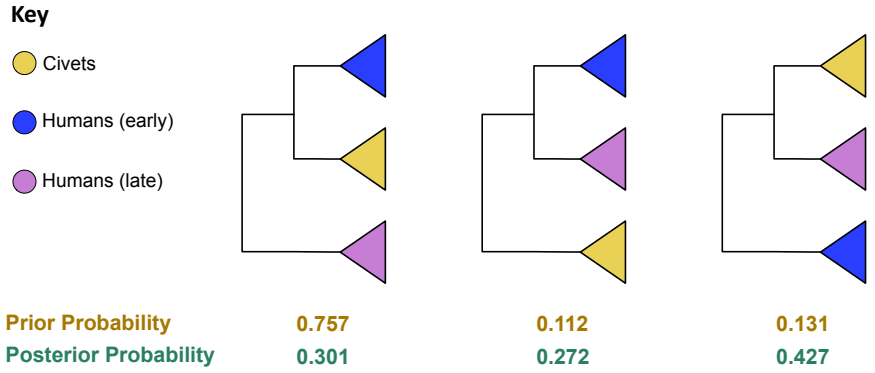


Figure 3.1: The evolutionary hypotheses for phylogenetic relationship of the epidemic SARS-CoV strains. These trees are mutually exclusively, i.e. the probabilities add up to 1, since the three groups are monophyletic. The prior and posterior probabilities were obtained by calculating the frequency of the three topologies among the the prior and posterior tree distributions that were estimated in BEAST without and with sequence data. The civet, early and late human clades in these trees comprises 15, 6, and 48 sequences respectively.

Epidemic strains

The SARS-CoV sequences sampled during the outbreak, 2003-2004, from both humans and civets, shared a pairwise identity of 99.6%. The low phylogenetic signal between the two populations resulted in an equivocal rooting of the MCC (maximum clade credibility) trees between the civets and the early human group. Except for the two human sequences that were known to be epidemiologically linked to a civet SARS-CoV sequence sampled from a restaurant in 2004 (Wang et al., 2005), all epidemic groups clustered monophyletically with strong support (posterior probability = 1.0). The BF test calculated on the prior and posterior odds for the three possible hypotheses in Figure 3.1 indicated weak support for the early human sequences as the most basal

group in the phylogeny (Table A.1), with BF score of 35.65 and 1.67 over the late human and civet groups respectively (Table A.1). The BF scores for the civets as an outgroup were comparatively lower against the late and early humans, at 21.29 and 0.60 (Table A.1). The low BF values associated with Tree 2 versus Tree 3 comparisons highlights the insufficient power to differentiate between the two hypotheses due to the high sequence similarity of the civet and human sequences. Consequently, no clear conclusions about the source of the human outbreak or the role of the civets during the epidemic can be drawn from these results.

	Early	Late	Civet
Early	-	35.65	**1.674**
Late	0.03	-	0.05
Civet	**0.60**	21.29	-

Table 3.1: The Bayes factor results for the different hypotheses outlined in Figure 3.1, which describes the three potential evolutionary relationships of the epidemic strains.

Divergence times that were estimated for the civet SARS-CoV (between June 2002 and April 2003) and the SARS epidemic strains (between April 2001 and December 2002) revealed the temporal sampling bias of the isolates collected during the outbreak. A bias towards the absence of early phase isolates from both humans and civets from the dataset.

Human and Bat Divergence

Recombination analysis with GARD and SBP verified the previously identified breakpoints in bat SARS-like CoV strains Rp3 and Rs672 (Lau et al., 2010; Hon et al., 2008; Yip et al., 2009). In addition, a group of closely related bat SARS-like CoV strains, called HKU3 (Figure 3.2), were also established as a recombinant lineage, generated along the same breakpoint. Accordingly, the phylogenetic analyses was performed separately on the flanking regions of the breakpoint, denoted by "part 1" and "part 2" respectively. The recombinant bat strains, Rp3 and Rs672, consistently clustered with

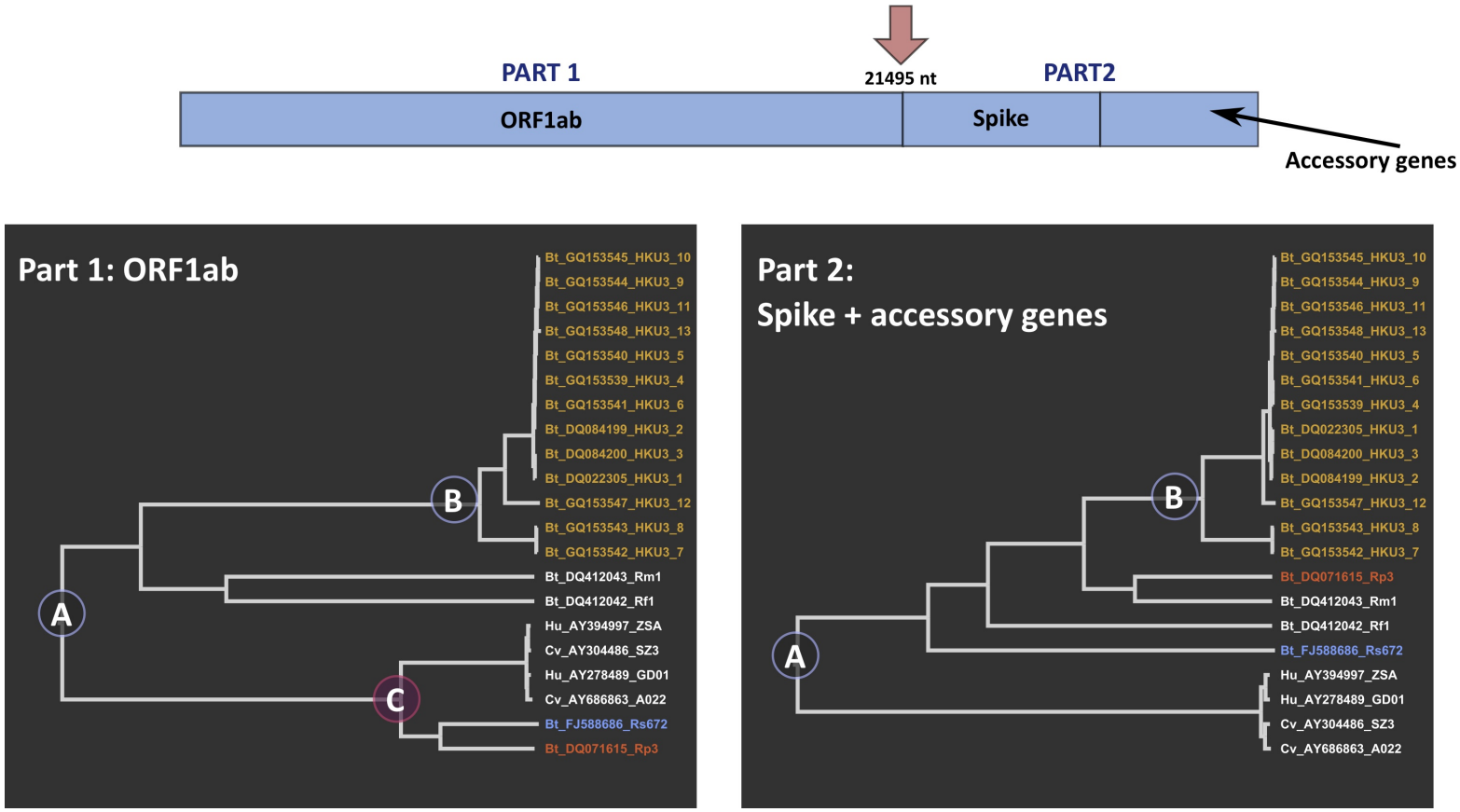


Figure 3.2: The maximum clade credibility (MCC) trees of the two regions of the SARS-CoV genome flanking the breakpoint (highlighted by the red arrow above) between ORF1ab and spike gene boundary. This includes all currently known bat SARS-CoV sequences, while the epidemic isolates (from humans and civets) are represented by a small subset. Bt = Bat, Cv = Civets and Hu = Humans. The labelled nodes A and B in both trees denote the same TMRCAs at the root and the bat SARS-CoV HKU3 clade. Node C in the ORF1ab tree indicates the TMRCA of human and recombinant bat SARS-CoV lineage (labelled in red and blue: Rp3 and Rs672 strains respectively).

Genome Region	Tree Model	Relaxed clock modell	A	B	C
part1	constant	gamma	1854.98 (1978.21-1729.53)	1986.38 (2001.60-1970.10)	1962.85 (1996.47-1925.39)
		uced	1929.48 (1995.51-1830.25)	1993.74 (2003.77-1977.78)	1985.60 (2001.55-1956.49)
		ucln	1881.50 (1999.89-1756.73)	1989.20 (2004.63-1973.22)	1971.08 (2002.66-1936.17)
	skyline	gamma	1781.77 (1878.10-1662.94)	1979.62 (1993.01-1965.16)	1942.78 (1976.22-1910.19)
		uced	1929.48 (1995.51-1830.25)	1983.24 (2002.85-1946.83)	1963.84 (1999.04-1908.97)
		ucln	1881.50 (1999.89-1756.73)	1978.51 (1992.22-1962.70)	1941.91 (1974.36-1905.67)
part2	constant	gamma	1923.73 (2002.85-1741.67)	1993.18 (2005.87-1987.43)	
		uced	1669.58 (1997.11-1142.52)	1954.20 (2003.98-1877.71)	
		ucln	1781.79 (2003.09-1361.91)	1966.51 (2004.88-1899.60)	
	skyline	gamma	1698.47 (1927.22-1347.87)	1963.24 (1998.08-1916.95)	
		uced	1743.75 (1968.36-1311.77)	1972.25 (2000.50-1920.12)	
		ucln	1686.24 (1937.39-1328.04)	1954.62 (1999.74-1985.77)	

Table 3.2: TMRCA estimates of the nodes A) currently sampled SARS-CoV from the three hosts (bats, civets and humans), B) the bat SARS-CoV HKU3 cluster and C) the human and recombinant bat lineage

the epidemic strains with a strong posterior support (0.99-1.0) in the part 1 regions, which is in line with previous findings that the epidemic strains have descended from a recombinant bat lineage (Figure 3.2). In the part 2 region, a different phylogenetic pattern was observed for Rp3 and Rs672 strains. While the clustering of Rp3 with other bat SARS-like CoV strains was associated with high posterior probability, unsurprisingly the inconsistent placement of Rs672 between independently sampled posterior distributions, was associated with weak statistical support (posterior probability ranging between 0.53-0.60). This observation most likely reflects the unsampled diversity of the bat SARS-like CoV strains in general, where the parental lineages of Rs672 were not represented in the dataset. The lack of related viruses in the dataset may also explain why the HKU3 group was not found to be phylogenetic incongruent between the two trees. The three recombinant lineages (Rp3, Rs672 and HKU3 strains) were likely to have been generated from different ancestral lineages, where the phylogenetic signal to determine the evolutionary history has become greatly confounded by multiple recombination events involving divergent strains. The recent divergence estimated for the human and recombinant bat lineage in the part 1 region by previous studies was specifically associated with the employment of the UCED clock and constant coalescent prior (Table 3.2). Although a BF test marginally preferred this combination of tree and clock model choice, the shape and scale parameters sampled for the UCGD clock model indicates that the rate variation among branches is unlikely to be exponentially distributed (shape=5.70 and 12.41 with constant and skyline coalescent priors respectively). Moreover, the mean coefficient of variation for UCGD and UCLN clocks range from 0.32-0.50 and 0.32-0.66 respectively, further illustrating that the UCED clock was not the most appropriate model to describe the part 1 sequence data, since there was no reasonable evidence for the substantial branch rate variation expected under the exponential distribution. The significantly younger dates estimated for the part 1 region from the UCED clock and constant coalescent model was linked to the interaction of the priors. Consequently, the fast evolutionary rates estimated for the deep branches in the phylogeny led to the compression of all the branches (Figure 3.3). The labelling

of the branches by median relative rate clearly demonstrated this effect in Figure 3.3, where all except the UCED clock and constant prior combination were represented by slight among branch rate variation (indicated by blue branches in Figure 3.3). To help visually compare the branch lengths in Figure 3.3, the scale bars have been adjusted to represent 30.0 years of evolution. A consistent pattern is observed when comparing the constant and skyline coalescent prior results, where shorter branch lengths are estimated when the constant population is employed (Figure 3.3). In contrast, the results from the skyline model were found to be similar for all three clock models. This suggests that the constant coalescent prior may not be appropriate model of demographic history for these SARS-CoV sequences. Given that the we have very limited samples of bat SARS-CoV, the constant coalescent prior is not likely to be a good model choice due to the restrictive assumptions it imposes on the unknown demographic history. These results strongly indicate that the estimated date of human and bat virus lineage of around 1998 was an artefact of the prior model choices employed by previous studies (Hon et al., 2008; Lau et al., 2010; Yip et al., 2009). Moreover, this highlights that the BF test, based on harmonic mean estimation (Newton and Raftery, 1994), currently implemented in the BEAST package, incorrectly selected this model combination. This is further supported by the mean molecular clock estimates of the human and bat lineage when the skyline coalescent prior was employed, which ranged from 1941-1963, with overlapping 95% highest posterior density (HPD) intervals for all clock models.

The particularly recent estimate of 1997-1998 inferred by previous studies was due to the calibration of node C (i.e. TMRCA of epidemic strains, see Figure 3.2) with a date derived from the spike gene analysis, the main antigen for SARS-CoV (Hon et al., 2008; Lau et al., 2010; Yip et al., 2009). This approach was undertaken under the assumption that antigenic-encoding regions contained more temporal signal due to the elevated rates of evolution in contrast to the polymerase-encoding ORF1ab region (i.e. part 1). The strong informative prior that was placed on this node, along with the use of UCED clock and constant growth models to infer the phylogeny for part 1, further compressed the date of the human and bat lineage. The molecular clock analyses under

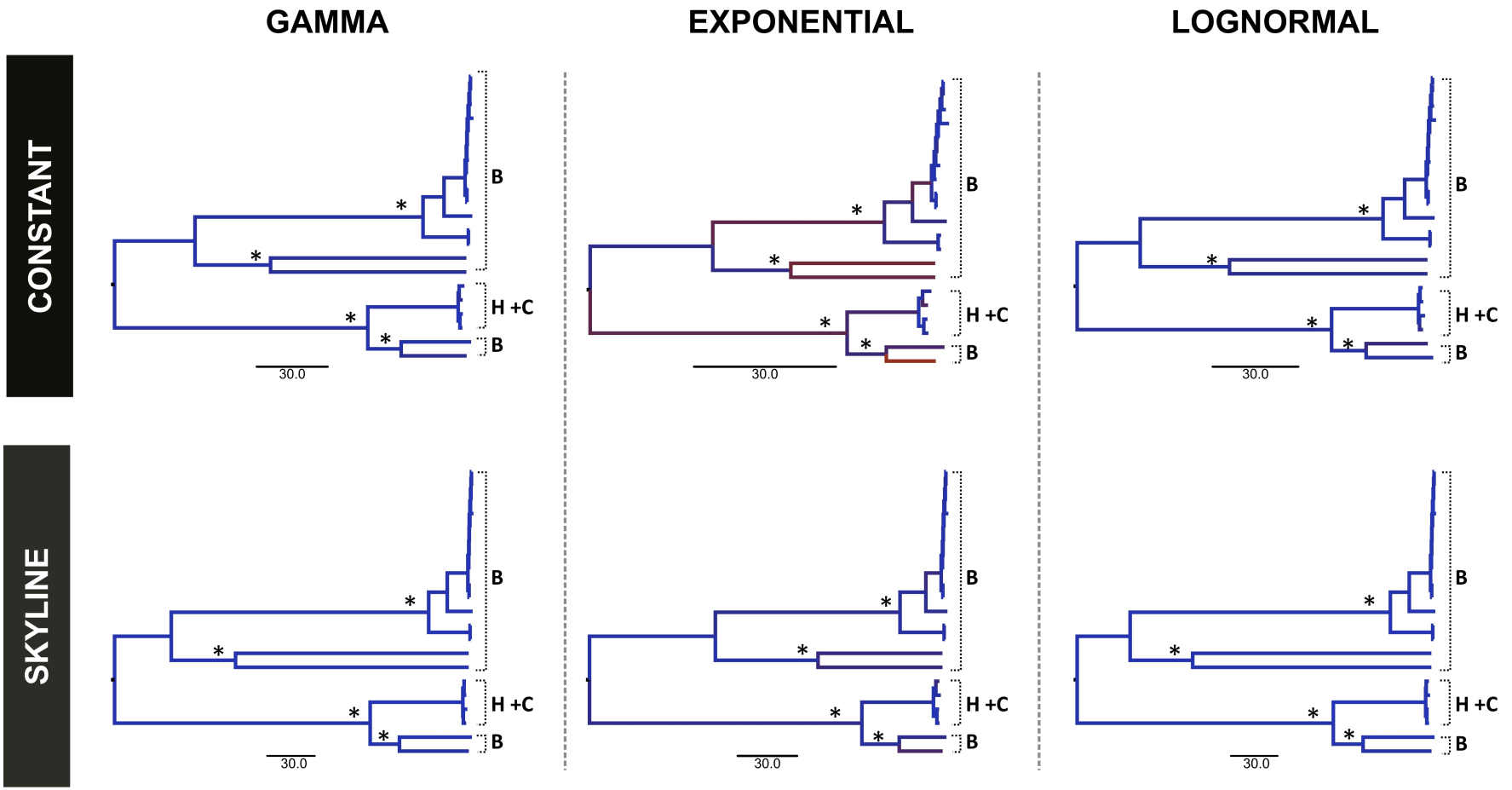


Figure 3.3: The MCC trees of part 1 (see Figure 3.1) estimated under different combination of relaxed clock and coalescent models on human-bat SARS-CoV dataset. The branches have been coloured by relative rates (same range across all trees). While the topology of the MCC trees do not vary between the different model choices, the combination of UCED clock and constant prior has resulted in dramatically contracting the branches (NB the scale bars have been scaled to represent 30.0 years for each tree). In the MCC tree estimated using UCED clock and constant coalescent prior has resulted in higher rates in the deeper branches, which is likely to explain the surprisingly recent dates of divergence between the bat ans human SARS-CoV lineage estimated by Hon et al. (2008) (see also Table 3.2). The branches with posterior support greater than 0.9 are marked by (*). B = Bats, H = Humans, and C = Civets.

both constant and skyline coalescent models estimated the mean TMRCA between 1941 and 1985 (Table 3.2). The fact that the 95% HPD intervals for these estimates coincided with a few years before the human outbreak (the lower HPD values ranged from 1996 to 2002), most likely reflected the uncertainty in the phylogeny, rather than being of any epidemiological significance. The likely under-sampling of the bat SARS-like CoV genetic diversity further suggests that additional bat viral isolates in the future will narrow this window of estimate of the human-bat SARS-CoV divergence.

3.4 Discussion

Understanding the emergence of SARS-CoV is crucial to assessing the risk of future cross-species transmission of related coronaviruses. To this end, both the role of the animal markets and the bat viral reservoir have been examined in a statistically robust framework as the immediate source of human SARS-CoV. The results highlight the challenges of studying novel appearances of zoonotic infections in human populations, where the absence of important epidemiological samples from the early stages of the outbreak can limit our view into the evolutionary origins of the virus. The epidemic strains, which include the human and civet SARS-CoV sequences, are clearly restricted in temporal sampling, where most have been isolated between 2003-2004, after the virus spread globally. The important note here is that these sequences can only provide limited information about the underlying process due to their high sequence similarity.

Additionally, recombination appears to be a pervasive force among this group of viruses, at least in the natural viral reservoir the bats. Recombination is widely observed in the *Coronaviridae* family, indicating that there is an inherent mechanism to allow genetic exchange and the opportunity for cells to co-infected with different viral strains is not lacking. However, whether recombination has played a significant part in the emergence of the human SARS-CoV strain is still not clear. While recent findings report two recombinant bat SARS-CoV strains being most closely related to the epidemic strains, it is unlikely that this lineage represents the immediate precursor strain of the SARS outbreak in 2002. The recent molecular clock estimates inferred by previous

studies for the human/bat SARS-CoV lineage appear to be biased by model choices and serves as a caution about prior interactions in Bayesian phylogenetic analyses (Holder and Lewis, 2003; Alfaro and Holder, 2006).

Given the results from this study, the currently sampled viruses from the animal markets are unlikely to be the direct source of the human SARS-CoV. However, since this is associated with weak support, it would be unwise to rule out the role of civets in the past and future emergence of related coronaviruses in humans. To understand the role of civets and determine the likelihood of SARS-like CoV re-emerging in the future, we need to increase our viral sequence samples from both early epidemiological cases and animal populations. While this study has highlighted some of the uncertainties surrounding the cross-species transmission of SARS-CoV, long-term surveillance of civets, bats and other ecologically linked species in the wild will help identify future emergent strains that are likely to pose a threat to humans.

The dates of divergence between the bats and human/civet SARS-CoV indicate that it is unlikely the bat viruses included in this investigation represent the immediate ancestors of the human outbreak. The more recent origin of the human/bat SARS-CoV lineage obtained by Hon et al. (2008) is entirely due to employment of UCED clock and constant coalescent prior. Although the datasets collated in this study differ slightly, the fact that we can recover the strikingly recent estimate with this particular combination of priors suggests the results obtained by Hon et al. (2008) are underestimated. In this study, the branch leading to the epidemic strains in part 1 represents 17-61 years of unsampled viral diversity from the bat reservoir. These estimates are compatible with the observation that the current sampled bat SARS-CoV cannot not bind the human receptor, ACE2, (Graham and Baric, 2010) indicating a much older divergence between the human and bat viral lineage. The lack of positive selection and adaptation in the human spike gene (Holmes and Rambaut, 2004), together with the vast undersampling of the bat SARS-CoV reservoir, suggests it is likely that there is a bat SARS-CoV lineage that can infect humans directly. In other words, the divergence between the recombinant bat SARS-CoV lineages discovered by Hon et al. (2008) and Yip et al.

(2009) and human/civet SARS lineage indicates that there may be non-recombinant bat viruses that share all the features of human SARS-CoV. The long period of unsampled diversity and the probable wide host-range of the epidemic strain (Martina et al., 2003; Li et al., 2006), could alternatively support the origin of the human outbreak in an yet unidentified animal reservoir.

The recent isolation of SARS-like CoV in bat species from Africa and Bulgaria demonstrates that our knowledge on the diversity and ecology of SARS-CoV and related coronaviruses is severely limited (Drexler et al., 2010; Tong et al., 2009). While presence of recombination and potential involvement of multiple host reservoirs makes the prospect of identifying emergent strains a challenging task, significant steps have been made in understanding coronaviruses in the wild. Although the role of civets may have been important in the emergence of SARS-CoV in humans, it is clear that the natural hosts, the bats, are likely to be the most important in determining the risk of future human outbreaks. Given that bats are also natural hosts for many other human pathogens (e.g. ebolaviruses and hantaviruses) (Leroy et al., 2005), including other coronaviruses (Cui et al., 2007; Drexler et al., 2010; Gouilh et al., 2011; Tang et al., 2006; Tong et al., 2009; Vijaykrishna et al., 2007), these group of species should receive particular attention when investigating emerging viral diseases.

4

The Origin and Evolution of the unique HCV circulating recombinant form 2k/1b

The manuscript form of this chapter has been published in the ***Journal of Virology, 86: 2212-2220*** as, "The origin and evolution of the unique HCV circulating recombinant form 2k/1b", (Raghwani et al., 2012).

This study includes some new sequence data from Amsterdam, with a large part collated from GenBank. The collection and sequencing of the new samples was performed by Xiomara V. Thomas, Sylvie M. Koekkoek, Janke Schinkel, Richard Molenkamp and Thijs J. van de Laar. Yutaka Takebe, Yasuhito Tanaka, and Masashi Mizokami provided helpful comments to the manuscript.

I performed the analyses, interpreted the results, and wrote the manuscript.

O. G. Pybus and A. Rambaut provided editorial and supervisory assistance.

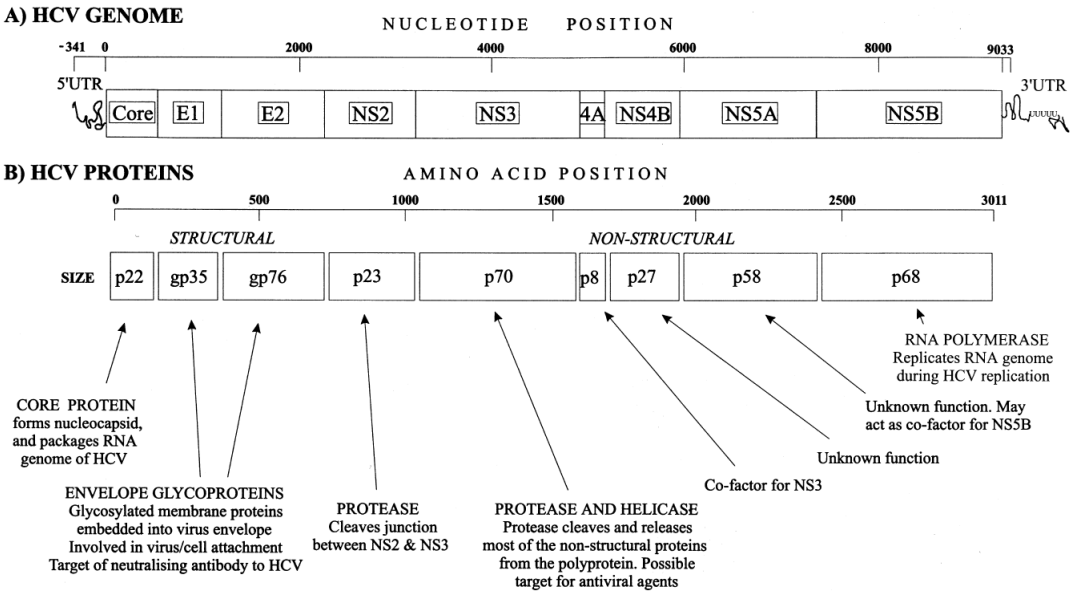
4.1 Introduction

In this chapter I investigate the only known circulating natural recombinant strain of the hepatitis C virus (HCV), 2k/1b (Kalinina et al., 2002). Since its initial discovery in St Petersburg, Russia in 1999, the strain has been isolated from several countries throughout Eurasia (Kalinina et al., 2002). It is formally classified as CRF01_1b2k Kuiken and Simmonds (2009), indicating that it is the only known circulating recombinant form (CRF) among HCV. Specifically, the strain has a 5' genome region that is most closely related to subtype 2k and a 3 genome region most closely related to the global epidemic subtype 1b, with a single recombination breakpoint located at genomic position 3175-3176 in the NS2 gene (see Figure 4.1) (Kalinina et al., 2002). There have been several other reports of inter- and intra-genotypic HCV recombinants in natural populations, although the evidence presented for recombination varies in strength; the weakest studies report only discordant genotyping results between genome regions (which could also result from co-infection) whereas the most convincing studies repeatedly sequence the same recombination breakpoint from independent extractions, thereby excluding the possibility of in vitro genetic exchange. Thus far there have been ten descriptions of HCV recombinant forms, although only in six cases have the breakpoints been sequenced (see Table 4.1) (Calado et al., 2011; Colina et al., 2004; Cristina and Colina, 2006; Kageyama et al., 2006; Lee et al., 2010; Legrand-Abravanel

Table 4.1: The currently known HCV recombinants.

Recombinant	5' Genotype	3' Genotype	Breakpoint Sequenced	Reference
2b/1b	2	1	NS3	Kageyama et al. (2006)
2i/6p	2	6	NS2/NS3	Noppornpanth et al. (2006)
2k/1b	2	1	NS2	Kalinina et al. (2002)
1b/1a	1	1	-	Moreno et al. (2009)
2k/5a	2	5	NS2/NS3	Legrand-Abravanel et al. (2007)
1a/1c	1	1	E1/E2	Cristina and Colina (2006)
2a/1a	2	1	-	Lee et al. (2010)
3a/1b	3	1	-	Lee et al. (2010)
2b/6w	2	6	NS2/NS3	Lee et al. (2010)
4d/4a	4	4	-	Calado et al. (2011)

Figure 4.1: A figure adapted from Simmonds (2001) to illustrate the organization of the HCV genome and the function of the encoding regions



et al., 2007; Noppornpanth et al., 2006).

Closer inspection of the reported breakpoint positions among HCV recombinants reveals a difference between inter- and intra-genotypic recombinants. Breakpoints in the intra-genotypic recombinants (1a/1c and 1b/1a) are located in the E1/E2 region, which encode the antigenic parts of the virus, while in the inter-genotypic recombinants (including CRF01_1b2k) the breakpoints are consistently found in the NS2-NS3 region (Cristina and Colina, 2006; Kageyama et al., 2006; Lee et al., 2010; Legrand-Abravanel et al., 2007; Moreno et al., 2009; Noppornpanth et al., 2006). Furthermore, naturally occurring inter-genotypic HCV recombinants have typically involved genotype 2 in the 5 genome region (Kageyama et al., 2006; Kalinina et al., 2002; Lee et al., 2010; Legrand-Abravanel et al., 2007; Noppornpanth et al., 2006) (Table 4.1), suggesting the genotype 2 may have some inherent property, either biological or ecological, that results in viable recombinant viruses.

As CRF01_1b2k is the only HCV recombinant with widespread dispersal, it has raised questions about the origins and dissemination of this unique CRF. Molecular clock and phylogenetic analyses have been useful in understanding and reconstructing

the epidemic history of various HCV strains, including subtypes 1a and 1b worldwide (Magiorkinis et al., 2009; Pybus et al., 2001) and subtype 1b in Japan (Tanaka et al., 2005). Similar analyses of HCV genotype 4 in Egypt have estimated the timescale of the large HCV epidemic in that country and have confirmed its iatrogenic cause (via inadvertent transmission or complications from medical treatment) (Pybus et al., 2003; Tanaka et al., 2004). Very little is known about the evolutionary history of HCV subtype 2k, most likely because of the lack of sequence data for the strain.

In order to address the lack of information about the epidemiological and transmission history of HCV CRF01_1b2k, in this chapter a comprehensive evolutionary analysis of all available viral genome sequences is conducted in a Bayesian phylogenetic framework (Drummond and Rambaut, 2007). Since previously published sequence data on CRF01_1b2k are limited, the sample size was increased by isolating and sequencing a panel of new recombinant isolates from patients of Russian and Georgian origin resident in Amsterdam (performed by X. V. Thomas and colleagues from the Academic Medical Center, Amsterdam).

By combining information from different genomic regions in a single analysis, this study provides the first estimate of the date of the recombination event that generated CRF01_1b2k. The date obtained considerably predates the discovery of the strain and requires a re-evaluation of the circumstances surrounding its formation. Furthermore, we estimate the CRFs past rate of transmission and its pattern of global geographic spread. A hierarchical Bayesian model was employed to jointly estimate the date of emergence of the CRF01_1b2k lineage from flanking regions of the genome. This approach results in greater precision of parameter estimates, thus improving on methods previously applied to dating the origins of HIV-1 CRFs, where each region was analysed separately (Ristic et al., 2011; Tee et al., 2009). The methods introduced here should serve as a model for future phylogenetic investigations of genetic-exchange events in RNA virus populations.

4.2 Methods

Identification of new HCV 2k/1b isolates from Amsterdam

In the course of a study of HCV-infected patients resident in Amsterdam (*Thijs et al, 2011* unpublished) it was found that HCV genotyping results from the 5UTR and NS5B regions were discordant for 6 (out of 200) patients. Of these, five were male and one female and their mean age was 34 (Table 4.2). The sequencing information provided by X. V. Thomas can be found in Appendix B.

In addition to new CRF01_1b2k sequences from Amsterdam, 6 recombinant isolates from an unpublished study were also included. These sequences were sampled from IDUs in Azerbaijan (Table 4.2). Most of the remaining CRF01_1b2k isolates come from a study of IDUs in Uzbekistan (Kurbanov et al., 2008a) and from a cohort study of HCV-positive patients from seven countries (Kurbanov et al., 2008b)); detailed demographic information on isolates from these two studies is provided here (Table 4.2). In the latter study, individuals infected with CRF01_1b2k all came from either Russia or Uzbekistan, and 6 of these patients were linked to high-risk groups, namely blood transfusion or intravenous drugs.

Collation of HCV sequence alignments

To investigate the evolutionary origin and spread of HCV CRF01_1b2k, a data set comprising all subtype 2k/1b (n=27), 2k (n=15) and 1b (n=71) isolates for which both core/E1 and NS5B sequences were available was collated from Genbank and the HCV Sequence Database ((Kuiken et al., 2005); Table B.1). This collection included the 6 newly sequenced isolates obtained from HCV patients in Amsterdam (see above). Alignments for both genome regions were constructed manually and each alignment contained exactly the same set of taxa. The date and sampling location of each sequence was obtained from the literature or via personal communication (Table 4.2).

Estimation of genome region-specific rates of evolution

The evolutionary rates of the core/E1 and NS5B regions used in this study could not be estimated directly from our data, because the sample size and range of sample dates were not large or wide enough. In line with previous studies of HCV epidemic history (Pybus et al., 2003), the substitution rates of the regions of interest were estimated from an independent data set with significant temporal information. Specifically, the data and analysis strategy of a recent study that reported rates of evolution for the HCV genome, which used an alignment-partition approach implemented in BEAST (Drummond and Rambaut, 2007) to estimate region-specific rates (Gray et al., 2011), were employed for both subtypes 1a and 1b. A codon-structured nucleotide substitution model (Shapiro et al., 2006a), a relaxed uncorrelated lognormal molecular clock (Drummond et al., 2006) and a Bayesian skyline coalescent model (Drummond et al., 2005) were applied to both subtype 1a and 1b whole genome alignments, from which two rate estimates were obtained for the sub-genomic regions, core/E1 and NS5B. The MCMC chains were run for 200 million generations and sampled sparsely to yield a posterior tree distribution based upon 10,000 estimates. For further analysis details, see Gray et al. (2011).

Phylogenetic analysis

Preliminary phylogenetic analyses were undertaken to confirm that CRF01_1b2k originated from a single recombination event. Neighbour-joining (NJ) trees of the core/E1 and NS5B datasets were estimated using PAUP* (Swofford, 2003) with a HKY85 nucleotide substitution model and a gamma-distributed among site rate heterogeneity.

Next, in order to directly test the hypothesis of a single recombinant origin, a Bayesian MCMC analysis of the core/E1 and NS5B datasets was performed in two ways: (i) the CRF01_1b2k isolates were constrained to be a monophyletic clade, and (ii) no phylogenetic constraints were imposed. The hypothesis of a single origin was then tested by performing a Bayes factor comparison of the marginal likelihoods (Newton and Raftery, 1994; Suchard et al., 2001) obtained from these two analyses. This

revealed an insignificant difference between the two competing hypotheses, thus a single recombination event was assumed in following analyses.

Molecular clock analysis

In order to estimate the date of the recombination event that formed CRF01_1b2k, separate datasets were assembled for the core/E1 and NS5B regions that contained the CRF isolates, plus all available closely-related parental subtype reference sequences (belonging to subtype 2k for the core/E1 region, and to subtype 1b for the NS5B region). A hierarchical phylogenetic model (Suchard et al., 2003) was used to combine information from both datasets and jointly estimate the time of the most recent common ancestor (TMRCA) of the CRF01_1b2k clade, whilst accounting for uncertainty in both genome regions. Specifically, separate phylogenies, molecular clock models and substitution models were estimated for the core/E1 and NS5B regions. The genome region-specific rates (estimated as described above) were used as prior distributions for the evolutionary rates for core/E1 and NS5B regions. For the NS5B region, the rates estimated from the subtype 1b were applied, while for core/E1 region the average of the 1a and 1b rates were used (because a subtype 2k-specific rate was not available).

For each pair of sampled phylogenies (core/E1 and NS5B) in the posterior distribution of the MCMC, three node dates were obtained (as labelled in Figure 4.3): (A) the joint TMRCA of the CRF clade, (B) the date of the parental node of the CRF clade in the core/E1 subtype 2k phylogeny, and (C) the date of the parental node of the CRF clade in the NS5B subtype 1b phylogeny. The former date, together with the more recent of the latter two dates, therefore define a time range during which the recombination event must have occurred (see Figure 4.3). The posterior distribution of this time range was then compiled by repeating the above procedure for each pair of phylogenies in the MCMC output. BEAST analysis model settings were the same as those outlined in the genome region-specific rates section above.

CRF01_1b2k transmission history

Further Bayesian MCMC phylogenetic analyses were performed solely on the CRF01_1b2k isolates in order to estimate the epidemic history and basic reproductive number, R_0, of the strain since its emergence. BEAST model settings were the same as those outlined above, except that different coalescent models were employed to reconstruct the transmission history of the CRF. Both the GMRF skyride (Minin et al., 2008) and exponential growth coalescent models were used.

4.3 Results

Estimation of genome region-specific rates of evolution

The rates of evolution of the core/E1 and NS5B genomic regions used in this study were estimated from subtype 1a and 1b whole genome data sets (see Methods) and are given in Table 4.3. Although there was some overlap in the highest posterior density (HPD) intervals for the rates between subtypes, estimated rates did show some variation among subtypes (Table 4.3).

Phylogenetic analysis

To establish the evolutionary origins of the HCV 2k/1b strain, the core/E1 (644 nt) and NS5B (741 nt) regions of 27 CRF 1b/2k, 15 subtype 2k and 71 subtype 1b isolates were analysed. The 2k/1b isolates were sampled between 1999 and 2007 from following locations: Ireland, Uzbekistan, Azerbaijan, Cyprus, Amsterdam, France and Russia (Table 4.2). Since the Bayes Factor (BF) test supported the hypothesis that the 2k/1b isolates were monophyletic, a single origin of the CRF was inferred (BF scores for the comparisons of monophyly and non-monophyly models were 0.31 for the core/E1 data set and -0.68 for the NS5B data set). Furthermore, the monophyletic origin of the CRF clade was supported in both genome regions when no phylogenetic constraints were imposed (not shown). However, monophyly of CRF 1b/2k isolates was not supported by a high posterior probability (0.67) in the NS5B MCC tree, most likely reflecting

Table 4.2: Epidemiological and sequence information of the CRF01_1b2k isolates used in this study

Isolate Name	Age (Years)	Gender	Location	Risk Factors	Country of Origin	NS5B	Core/E1	NS2	Reference
1b2k_AZ_01AZ051_2000	34	M	Azerbaijan	IDU	Azerbaijan	FJ435529	FJ435462	-	Unpublished
1b2k_AZ_01AZ082_2000	38	M	Azerbaijan	IDU	Azerbaijan	FJ435544	FJ435480	-	Unpublished
1b2k_AZ_02AZ105_2001	32	M	Azerbaijan	IDU	Azerbaijan	FJ435550	FJ435490	-	Unpublished
1b2k_AZ_02AZ114_2001	41	M	Azerbaijan	IDU	Azerbaijan	FJ435556	FJ435497	-	Unpublished
1b2k_AZ_02AZ129_2001	30	M	Azerbaijan	IDU	Azerbaijan	FJ435564	FJ435505	-	Unpublished
1b2k_AZ_02AZ139_2001	33	M	Azerbaijan	IDU	Azerbaijan	FJ435572	FJ435514	-	Unpublished
1b2k_CY_CYHCV037_2005	-	-	Cyprus	ST/IDU	Georgia	EU684614	EU684686	-	(7)
1b2k_CY_CYHCV093_2007	-	-	Cyprus	ST/IDU	Georgia	EU684649	EU684728	-	(7)
1b2k_AM_P077_2006	35	M	Amsterdam	IDU	Russia	JF949902	JF949908	Confirmed	This Study
1b2k_AM_P079_2006	42	M	Amsterdam	-	Georgia	JF949897	JF949903	Confirmed	This Study
1b2k_AM_P108_2007	35	M	Amsterdam	IDU	Georgia	JF949898	JF949904	Confirmed	This Study
1b2k_AM_P135_2005	35	F	Amsterdam	-	Georgia	JF949899	JF949905	Confirmed	This Study
1b2k_AM_P159_2007	39	M	Amsterdam	-	Georgia	JF949900	JF949906	Confirmed	This Study
1b2k_AM_P179_2000	21	M	Amsterdam	IDU	Georgia	JF949901	JF949907	Confirmed	This Study
1b2k_FR_M21_2007	30	M	France	IDU	Georgia	FJ821465	FJ821465	Confirmed	(Morel et al., 2010)
1b2k_IE_HC9A99966_2006	-	-	Ireland	-	Russia	AB327058	AB327018	Confirmed	(Moreno et al., 2009)
1b2k_RU_747_1999	-	-	Russia	IDU	Russia	AF388411	AY070214	Confirmed	(Kalinina et al., 2001, 2002)
1b2k_RU_796_1999	-	-	Russia	-	Russia	AF388412	AY070215	Confirmed	(Kalinina et al., 2001, 2002)
1b2k_RU_ALT30_2000	34	F	Russia	-	Russia	AB327055	AB327015	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_HIA1002_2003	-	-	Russia	-	Russia	DQ001221	AB327011	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_KNG318_2002	30	M	Russia	IDU	Russia	AY764172	AB327010	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_KNG327_2002	25	M	Russia	-	Russia	AY764176	AB327012	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_N687_1999	-	-	Russia	-	Russia	AY587845	AY587845	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_PSA108_2005	37	M	Russia	-	Russia	AB327053	AB327013	Confirmed	(Kurbanov et al., 2008b)
1b2k_RU_PSA62_2005	50	M	Russia	BT	Russia	AB327054	AB327014	Confirmed	(Kurbanov et al., 2008b)
1b2k_UZ_AZ15	22	M	Uzbekistan	BT/IDU	Uzbekistan	AB327056	AB327016	Confirmed	(Kurbanov et al., 2008a))
1b2k_UZ_UZIDU19_2006	-	-	Uzbekistan	IDU	Uzbekistan	AB327120	AB327122	Confirmed	(Kurbanov et al., 2008a)

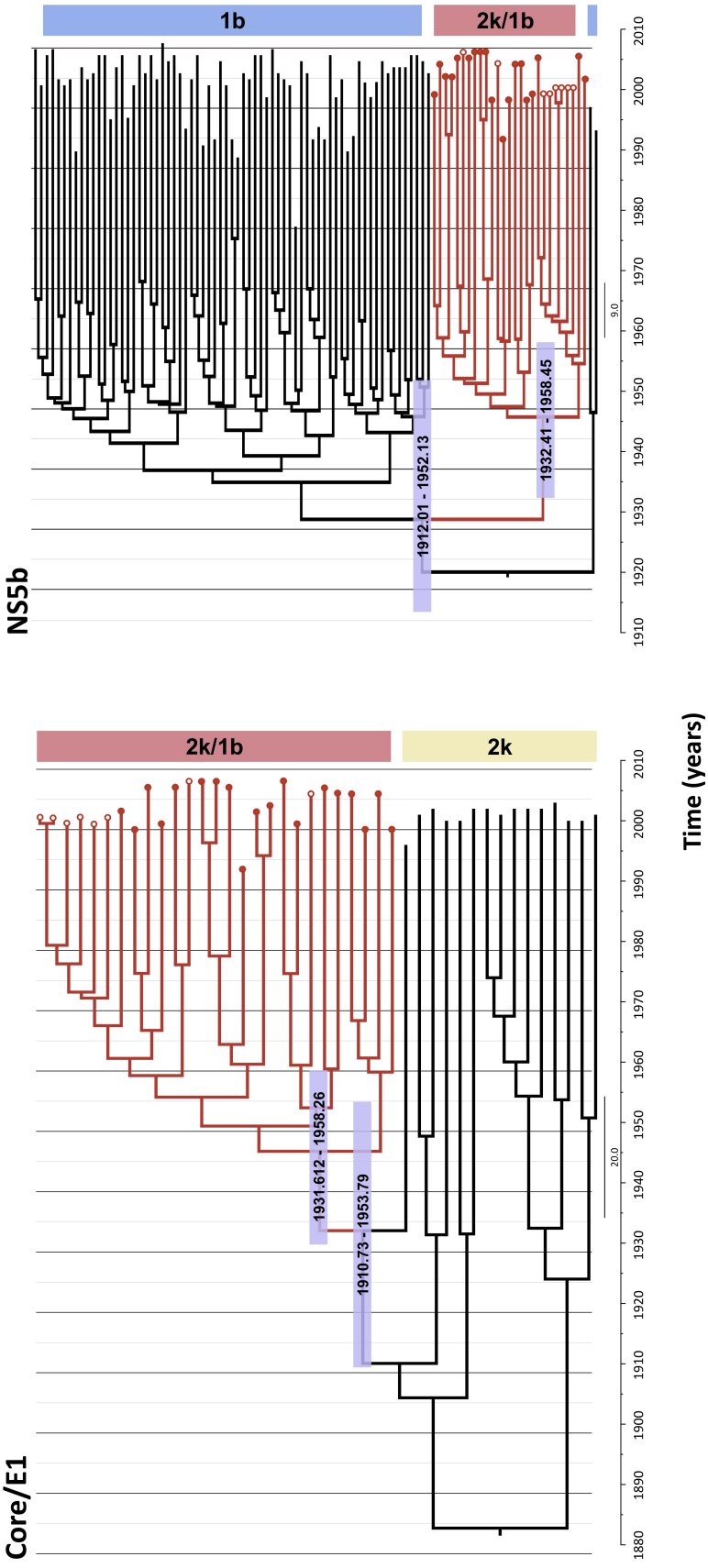


Figure 4.2: The maximum clade credibility trees of CRF01_1b2k and its parental subtypes, estimated from (a) the Core/E1 alignment and (b) the NS5B alignment. The horizontal bars indicate the dating estimates for two nodes: the common ancestor of the CRF clade (1931/32 to 1958) and the common ancestor of the CRF and the most closely related parental strain (1910/12 to 1952/53). The filled red circles indicate the 2k/1b isolates that were confirmed as being recombinant by sequencing of the breakpoint in the NS2 region. Open circles indicate those isolates for which NS2 sequences were not available.

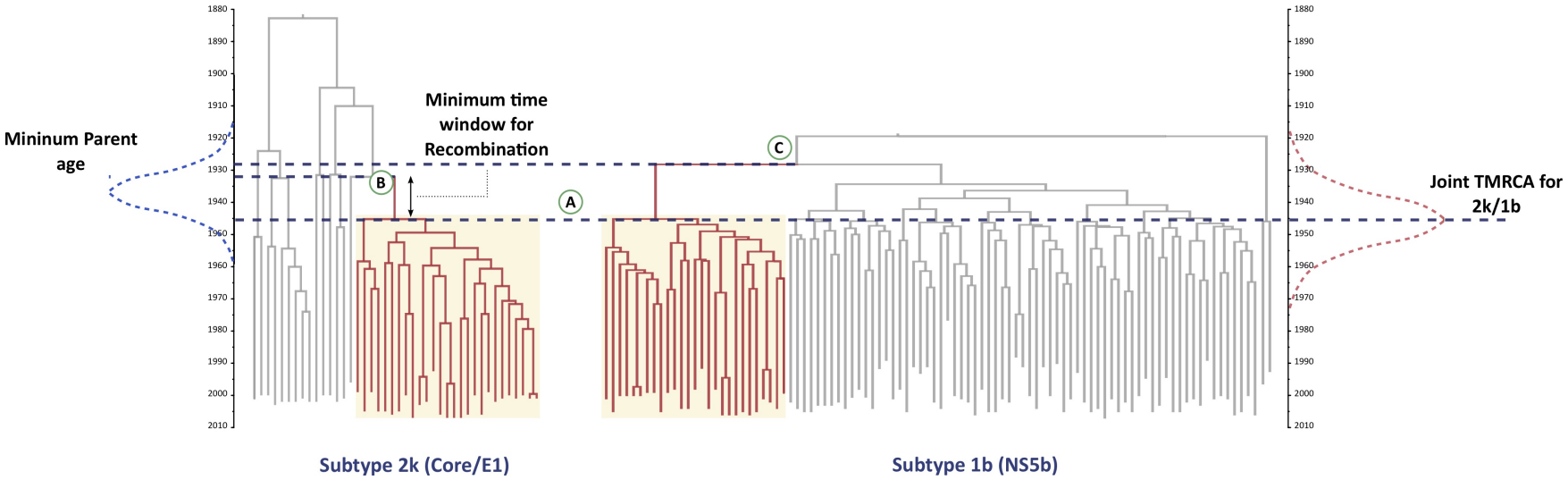


Figure 4.3: An illustration to explain the joint phylogenetic method that we employed to estimate the date of the recombination event that generated CRF01_1b2k. The left axis depicts the time of the ancestral nodes (B in core/E1 region, and C in NS5B region) and the right axis shows the time of the recombination event that generated CRF01_1b2k. Both the joint CRF node A, estimated from jointly from core/E1 and NS5B region, and the youngest parental node (either B or C) were use were used to calculate the minimum time window for the recombination event to have occurred.

Table 4.3: Evolutionary rate estimates of the genomic regions used in this study, obtained from HCV subtype 1a and 1b whole genomes using a genomic partition model (see Methods). The numbers inside parentheses represents the 95% highest posterior density credibility interval.

Genomic Region	HCV subtype 1a (subs/site/year x 10^{-3})	HCV subtype 1b (subs/site/year x 10^{-3})	Average Rate (subs/site/year x 10^{-3})
Core/E1	1.75 (1.41, 2.10)	1.36 (1.01, 1.76)	1.56 (1.21, 1.93)
NS5B	0.89 (0.71, 1.07)	0.91 (0.68, 1.14)	0.90 (0.70, 1.11)

the uncertainty associated with the star-like phylogeny and relatively short sequence length. However, the recombinant nature of many of the isolates in this study were confirmed by direct observation of the breakpoint in the NS2 gene region (confirmed recombinants are represented by filled circles in Figure 4.2).

Molecular clock analysis

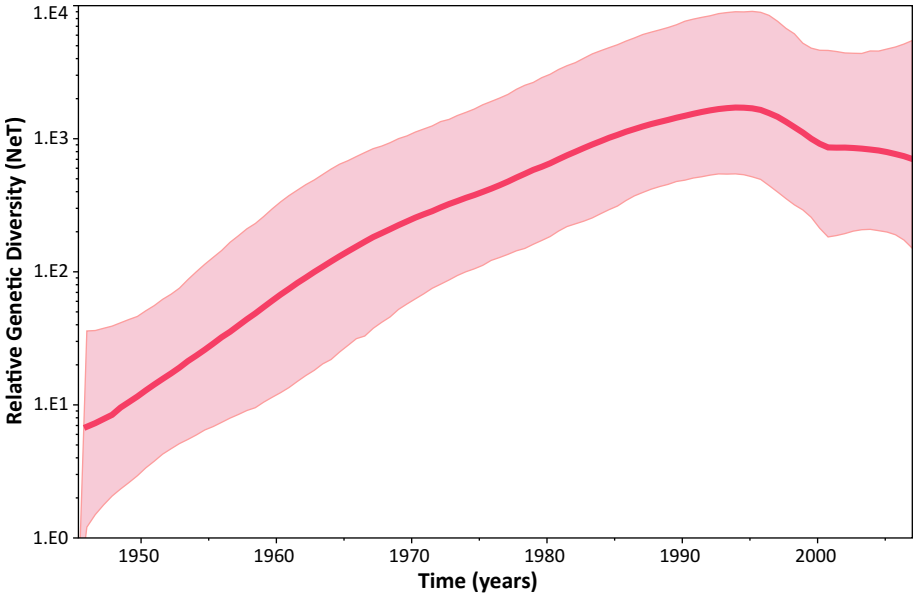
Figure 4.2 and Table 4.4 provide the estimated TMRCAs of the CRF clade obtained using the hierarchical phylogenetic and molecular clock analysis. The estimates obtained separately from the core/E1 and NS5B data sets were in close agreement and exhibit overlapping HPD intervals. These estimates also agree with the joint estimate of TMRCA of the CRF (node A) which was 1946 (1932-1959; Table 4.4). The date of the most recent common ancestor of the CRF clade and its most closely-related parental isolate was also estimated (Figure 4.3). These estimates were again similar for

Table 4.4: Estimates of the TMRCA of the CRF clade obtained from the separate genome regions, and from the joint (hierarchical) phylogenetic analysis. The numbers inside parentheses represents the 95% highest posterior density credibility interval.

TMRCA[a]	Core/E1	NS5B	Joint (Hierarchical) estimate
CRF (A)	1945.2 (1931.6,1958.3)	1945.8 (1932.4,1958.5)	1946.0 (1932.5, 1959.0)
CRF+2k (B)	1932.0 (1909.7,1952.8)	-	-
CRF+1b (C)	-	1933.1 (1912.0,1952.1)	-

[a]The letter in the brackets refers to the node labels highlighted in Figure 4.3

Figure 4.4: The estimated Bayesian skyride plot of CRF01_1b2k. The vertical axis represents the product of viral generation time T and the effective number of infections (N_e).



both genome regions, at around 1933 (Table 4.4). Lastly, by comparing the HPDs of the node A date with those of the more recent of the two parental nodes (either node B or C), a 95% HPD interval was obtained for the time of the recombination event that generated the CRF, which was between 1923 and 1956.

CRF01_1b2k transmission history

Figure 4.4 shows the epidemic history of the CRF01_1b2k estimated with the Bayesian skyride method. The skyride plot indicated an approximately constant exponential growth since the emergence of the CRF01_1b2k lineage (Figure 4.4) until the mid-1990s, after which the effective population size declined or stabilised (either could be plausible given the size of the credible region of the estimate). This decrease/stabilisation immediately followed the advent of screening for HCV in blood donors, which greatly reduced the risk of HCV infection via blood transfusion (Lemon and Brown, 1995; van der Poel, 1999; Schreiber et al., 1996).

To ascertain the exponential growth rate (r) of the HCV recombinant, the CRF01_1b2k data set was also analysed using an exponential growth coalescent model. The

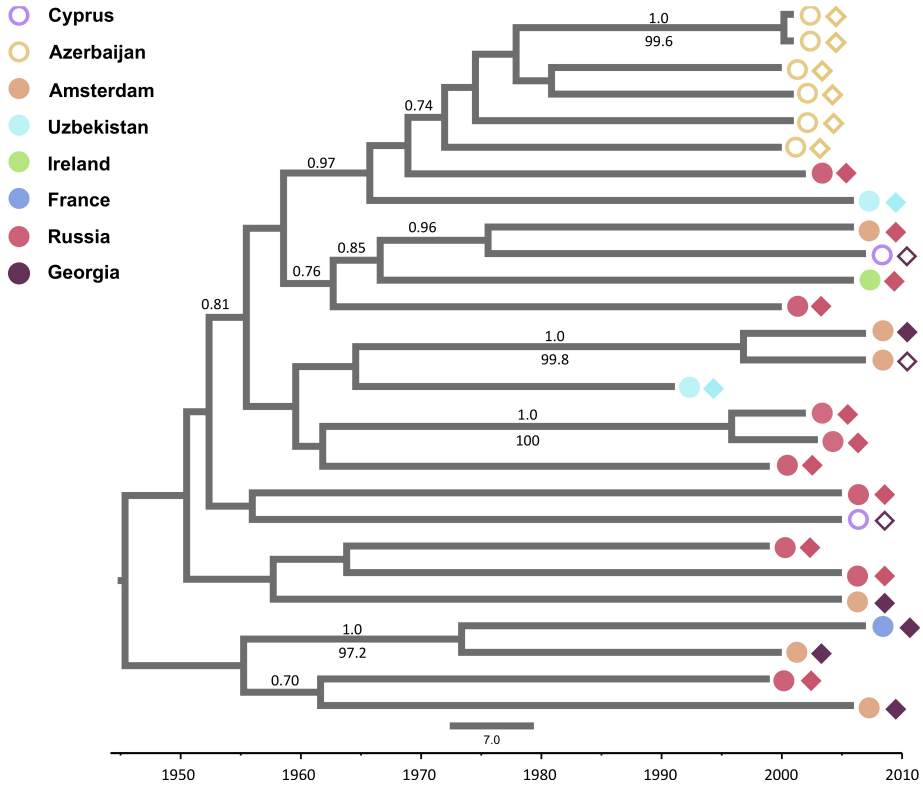


Figure 4.5: A molecular clock phylogeny of the CRF01_1b2k clade, estimated using a relaxed uncorrelated lognormal clock model and the SDR06 nucleotide substitution model (see Methods). Sequences are colour-labelled according to the country of origin (diamonds) and sampling (circles). All isolates were found to have an epidemiological link to Russia or the former Soviet Union; see Table 4.2). Filled circles/diamonds indicate 2k/1b isolates that were confirmed as being recombinant by sequencing of the breakpoint in the NS2 region. Open circles/diamonds indicate those isolates for which NS2 sequences were not available. The numbers above the branches show the posterior probability of nodes in the MCC tree; number below the branches represent bootstrap support values using maximum likelihood.

growth rate was estimated at 0.116 year^{-1} (95% HPD interval: 0.079 - 0.159), which subsequently used to calculate R_0 values under a plausible range of average durations of infectiousness (D) with the following equation: $R_0 = rD + 1$ (Pybus et al., 2001). Both the estimated growth rate (r=0.1 year^{-1}) and the estimated R_0 values (R_0 2-4) are compatible with a number of equivalent estimates for other HCV subtypes, including those obtained from IDU risk groups (Pybus et al., 2001, 2003, 2005; Tanaka et al., 2005).

The maximum clade credibility (MCC) tree of the CRF01_1b2k clade (Figure 4.5),

when combined with all available epidemiological information (Table 4.2), indicated a clear pattern of phylogenetic clustering according to the geographic location and risk factor of each patient. For example, where these details were readily available, 14 out of 15 isolates were associated with IDU, while only one was isolated from a patient with a history of blood transfusion (Table 4.3). These observations further support the view that CRF01_1b2k transmission is strongly linked with the IDU transmission route. Although it is not possible to reconstruct the location of the common ancestor of the CRF lineage with any certainty, all of the isolates included in this study are from (or have an epidemiological link to) the former Soviet Union, and the oldest strain was sampled in Russia in 1999. A well-supported cluster of strains from Azerbaijan originated around 1970 and therefore the CRF has likely circulated there since that time. Hence it appears that CRF01_1b2k disseminated throughout the former Soviet Union before its dissolution in the late 1980s, and for some time before the discovery of the CRF in St Petersburg in 1999. However, further samples and epidemiological data are required to study the global spread of CRF01_1b2k in more detail.

4.4 Discussion

As the only known HCV recombinant in widespread circulation, the existence and emergence of CRF01_1b2k presents an interesting question in HCV epidemiology and evolution. Investigating its evolutionary origins and transmission history helps to understand the circumstances that led to its unique properties. In contrast to HIV, which has 49 known CRFs and a much greater number of unique recombinant forms (Leitner et al., 2005), recombination typically contributes little to the generation and maintenance of HCV genetic diversity. Given that HCV has a higher global prevalence than HIV, and thus has higher likelihood of dual infections occurring in comparison, it is unlikely that epidemiological factors are restricting the opportunities for HCV to generate CRFs. Mixed infections with divergent HCV strains have been reported for many different populations, and are noted to be prevalent amongst high-risk group, particularly IDU and some haemophiliacs (Blackard and Sherman, 2007; Bowden et al., 2005;

Herring et al., 2004; Qian et al., 2000; Schröter et al., 2003).

Since the opportunities for HCV recombination are not limited, it is more likely that fundamental molecular and evolutionary differences between the HIV and HCV explain why HIV has many CRFs and HCV has few. These could include differences in the rate of template switching, or differences in genomic or immunological constraints, such that HCV recombinants have on average lower fitness than HIV recombinants, and therefore are rarely transmitted (Worobey and Holmes, 1999). Although both viruses are associated with chronic infections, unlike HIV, HCV can be spontaneously cleared by the host. This may explain in part the differences in the number of recombinants between HIV and HCV, where partial protective immunity in the latter reduces that chance of in vivo recombination of HCV strains (Aitken et al., 2008; van de Laar et al., 2009; Osburn et al., 2010). However, the high rate of mixed infections observed suggests that this is likely at best to play a minor role in HCV recombination. The low frequency of HCV recombinants is more likely to reflect mechanistic constraints at viral replication. There is evidence that template switching in HCV is especially rare and that the replication complex is typically encoded on the same genomic strand that it will replicate and transcribe (Appel et al., 2005). It is also interesting that when replication complexes are exchanged between different genotypes, the replication efficiency is substantially reduced (Herlihy et al., 2008). The pseudo-diploidy of the HIV genome certainly increases the likelihood of recombination occurring due to the ability of the virus to package two RNA templates (Hu and Temin, 1990), while secondary RNA structure in the HCV genome may limit the production of viable hybrid HCV viruses (Simmonds and Smith, 1999; Tuplin et al., 2002).

Our study provides the first estimates of the date of the recombination event that generated CRF01_1b2k. We estimated time of origin of CRF01_1b2k to be between 1923-1956, which is not much later than the origin and global spread of the parental subtype 1b (Magiorkinis et al., 2009). This date is significantly earlier than expected, since it was previously thought the genesis of CRF01 was linked to the dramatic increase in IDU behaviour following the break-up of the former Soviet Union. The involvement of

subtype 1b in the recombinant is unsurprising, as it is one of the most prevalent subtypes worldwide. However, to fully appreciate the origin of CRF01_1b2k, we need to consider the phylogeography of both parental subtypes and of the recombinant lineage itself. Genotype 2 harbours considerable genetic diversity, especially in West Africa, which is where the genotype is thought to have originated (Markov et al., 2009). Although the small number of subtype 2k isolates sampled to date likely underestimate the true extent of its distribution, such viruses have been isolated from Martinique and Madagascar, implicating the historical trans-Atlantic slave trade in the dissemination of the virus from West Africa (Markov et al., 2009).

The current distribution of subtype 2k is associated with francophone regions and former-Soviet Union countries. In contrast, CRF01_1b2k is more spatially limited, with all isolates being directly or indirectly linked to the former Soviet Union. Therefore it seems most likely that CRF01_1b2k was generated in the former USSR. The non-recombinant subtype 2k isolates that are most closely related to CRF01_1b2k are from Moldova and Azerbaijan (Figure 4.2). An equivalent analysis of subtype 1b viruses provides no reliable phylogenetic linkage, due to the low phylogenetic resolution of the NS5B data set.

Interestingly, the estimated date of CRF01_1b2k coincides with the time when the former Soviet Union founded the first national blood transfusion service in the 1920s. As an early leader in transfusion technology, a large-scale network comprising of blood transfusion centres and research institute across the republic. The Soviets also adopted blood storage and preservation techniques, leading to the establishment of more than 60 primary and 500 subsidiary blood storage centres by the mid-1930s, which shipped blood across the entire Soviet Union (Starr, 1999). During the Second World War these networks were swiftly re-adapted to support the front line; in Moscow alone around 2000 blood donations were given per day (Huestis, 2002; Starr, 1999). The impressive scale of the blood service in the USSR is likely to have favoured HCV transmission by increasing the efficiency and geographic range of CRF01_1b2k dissemination. Whether specific medical practices at this time increased the probability of mixed viral infec-

tions remains unknown. It is interesting to note that Bogdanov himself was fascinated by the ideological interpretation of blood-sharing and frequently practiced what he called physiological collectivism the exchange of blood with others through mutual transfusions (Starr, 1999).

As well as providing a credible hypothesis for the origin of CRF01_1b2k, via unscreened blood transfusions, we can look to historical circumstances to explain how subtype 2k, or the CRF itself, arrived in the Soviet Union from West Africa or the Caribbean. Migration from Africa to the USSR occurred during the late 1950s and 70s as result of alliances forged by the Soviet government with newly independent African states such as Ghana and Angola (Matusevich, 2009). However, according to our estimates, these connections are too late to have contributed to the emergence of CRF01_1b2k. Although we cannot reject the hypothesis that the CRF was formed in West Africa and subsequently travelled to the Soviet Union, our results are more consistent with the recombination event occurring in the Soviet Union. This uncertainty highlights the need for more samples, especially subtype 2k viruses from African and former Soviet Union locations.

The epidemic history CRF01_1b2k (Figure 4.4) since its emergence is similar to that estimated for other epidemic subtype of HCV (e.g. (Magiorkinis et al., 2009)). The growth in CRF01_1b2k effective population sizes coincides with a substantial increase in blood transfusion, including during the Second World War, and with the subsequent rise in intravenous drug usage. CRF01_1b2k transmission seems to have slowed or stabilised since the early 1990s, coinciding with the onset of the anti-HCV screening of donors. In the absence of any data to the contrary, the transmission of this recombinant, and its spread from the former Soviet Union, reflects the peculiar epidemiological properties of the risk groups it has been associated with rather than any intrinsic properties of the virus.

We demonstrate the practicality and benefits of jointly estimating parameters of interests when analysing multi-partite sequence data that results from genetic exchange. This approach yields more accurate parameter estimates in comparison to when ge-

nomic regions are analysed separately (e.g (Lam et al., 2008; Tee et al., 2009)), by jointly incorporating the phylogenetic information and uncertainty of different genomic regions.

This study has made significant steps in understanding the epidemic history and spread of the unique circulating HCV recombinant 2k/1b. Most significantly, we show that this strain originated many decades before the post-Soviet rise in injection behaviour with which it is currently associated. Based on the date of its origin and its molecular epidemiology, there are reasonable grounds to suppose that the Soviet Unions revolutionary blood service was instrumental in the CRFs early generation and continental-scale spread. This infrastructure may have facilitated the pan-Eurasian spread of other parenterally-transmitted blood-borne infections and this is an interesting question for future research.

5

Dating the Divergence of Influenza Viruses

5.1 Introduction

While viral emergence is often associated with recent cross-species transmissions (Wolfe et al., 2007), e.g. SARS-CoV, there are several fast-evolving human RNA viruses that have much older origins, HCV and measles (Simmonds, 2001). To understand the short-term evolutionary dynamics of current human viruses, including where they come from, we need to look at their long-term evolutionary association with humans. Therefore, in this chapter I introduce a hierarchical analytical framework to estimate the ancient divergence dates of human RNA viruses. Inferring ancient historical events from molecular data is a very controversial issue, not least because of the progressive difficulty in estimating the true number of substitutions over long evolutionary distances due to increasing occurrence of multiple substitutions at each site (Brown et al., 1979, 1982; Arbogast et al., 2002; Sullivan and Joyce, 2005). This effect known as saturation of sites leads to underestimates of divergence dates at increasingly longer timescales, where eventually the evolutionary change becomes indiscernible (Brown et al., 1982; Arbogast et al., 2002; Sullivan and Joyce, 2005). This problem is expected to be exacerbated among acute-infecting RNA viruses with their extremely fast evolutionary rates.

Therefore, while there are methods to correct for multiple hits in comparisons of sequence data, it is debatable whether at certain tree depths such methods can ever reliably recover the true number of substitutions, thus limiting how far back we can infer the evolutionary timescale. This suggests that substitutional models are simplistic, which assumes that evolution is a relatively homogeneous process over both short and long evolutionary timescales. This bound on estimating divergence time varies with sequence data type, since the nucleotides evolve more quickly than amino acid sequences largely due to the presence of third codon positions, which are mostly synonymous and functionally unconstrained. In addition, with evolutionary change encoded by just 4 characters, rather than 20 with amino acids, over time nucleotide sequences will approach saturation more rapidly.

Ancient viral divergences of fast-evolving RNA viruses may represent cross-species transmission events that have taken place considerably early in the human history. This could have occurred during the agricultural revolution, when human populations were growing and had increased contact with live animals, around 3,000 to 8,000 years ago (Pearce-Duvet, 2006; Diamond, 2002; Dobson and Carper, 1996). Alternatively, the ancient viral divergence could indicate that we have inherited the RNA virus pathogen from our last common ancestor with primates (Sharp, 2002). Therefore, the timescale of ancient viral divergences could range from thousands to millions of years. However, in spite of the presumed long-term association of some RNA viruses with their human hosts, it is not clear whether we should expect ancient viral divergences to have occurred over similar timescales as estimated for slow-evolving DNA viruses (Holmes, 2003b). This difference may be explained by the life-histories of human RNA viruses that have infected human populations throughout their history. Acute-infecting viruses require sufficiently large host populations to sustain long-term transmission (Cleaveland et al., 2001, 2007). For human populations, this only occurred once agricultural development was established, indicating that the origins of current acute-infecting RNA viruses may be at most few thousands of years ago. While evidence from endogenous virus elements (EVE) that have integrated into our genomes, as well other primates and mammals, indicates that pre-agricultural humans were confronted by RNA viruses, it is very likely that these pathogens were associated with persistent, rather than acute, infections (Cleaveland et al., 2001, 2007; Diamond, 2002; Horie et al., 2010; Katzourakis and Gifford, 2010).

Some current RNA viruses, which have been circulating in the human population over number of years, are characterised by very limited standing genetic diversity, e.g. seasonal influenza and dengue viruses. This could provide an alternative explanation why old divergences events among RNA viruses are estimated with surprisingly recent dates, thus conflicting with epidemiological and historical data (Holmes, 2003b). The high extinction rates associated with some RNA viruses suggests that the time of the most common recent ancestor (TMRCA) is typically young, often 1-3 years, compared

to the time of original cross-species transmission event (Holmes, 2003b). Therefore, this could mean that early lineages are unlikely to be sampled in the present-day, which would lead to greatly biased estimates for ancient divergences of human RNA pathogens. This hypothesis is linked to how the virus infect new susceptibles in the population, which will be determined by the level of cross-protective immunity among previously infected individuals. This could occur by antigenic evolution or by importing new viruses from zoonotic reservoirs. So while high extinction rates may be true for antigenically variable RNA viruses that are associated with partial immunity in the host, it is unlikely to hold for viruses that elicit strong cross-protective immunity (e.g. measles) or cause chronic infections, such as HIV and HCV.

In this chapter, I develop an analytical framework to estimate the divergence date of all influenza viruses. This investigation also includes evaluating the effect of current substitutional models to accurately infer dates of divergence over short and long evolutionary timescales. To overcome the difficulty of multiple substitutions, 1) I have used the polymerase genes to infer divergence times, since they are considered be the most conserved encoding regions of the viral genome (Rambaut et al., 2008; Webster et al., 1992), and 2) substitution models are optimised to utilise the maximum information for the relevant tree depth. To reduce the issue of multiple hosts and short-term association, individual clades that either emerged recently in humans or are currently circulating in avian and swine hosts, are analysed separately before inferring the deeper divergences of the influenza tree. In addition, a hierarchical Bayesian model (Suchard et al., 2003) is applied to improve the precision of the MRCA estimates by combining the information from the polymerase genes. without requiring them to have the same tree.

5.2 Methods

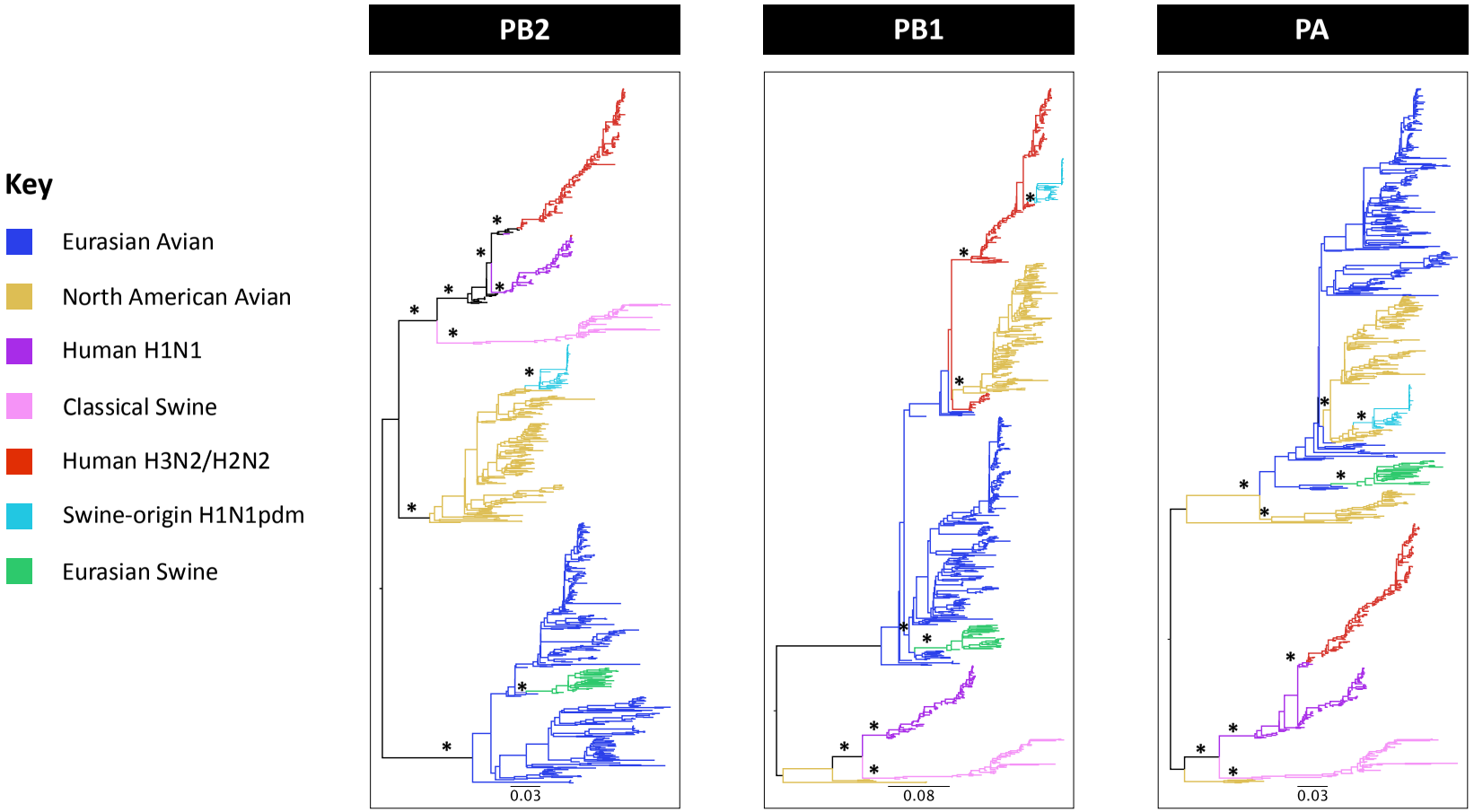
Short-term divergences

Influenza A viruses

Whole genome sequences for influenza A viruses isolated from the three hosts (avian, swine and humans) were downloaded from GenBank. Using the following selection criteria of one subtype per location per host per year (performed by Samantha J. Lycett), a representative sample of 891 viral isolates was collated for analysis. An alignment for each segment was assembled using the multiple alignment program, MUSCLE (Edgar, 2004) and subsequently verified by eye. To deduce the phylogenetic structure of the polymerase genes, a maximum-likelihood tree with 1000 bootstraps was constructed in RaxML (Stamatakis et al., 2005); using GTR nucleotide substitution model with gamma distributed rates (see Figure 5.1). This helped to identify the major clades of influenza A viruses, which included: Eurasian avian, North American avian, Classical swine H1N1, Human H3N2, Human H2N2, Human H1N1 (1918-), European swine H1N1, and swine-origin pandemic H1N1. Subsequently, each clade was then considered separately to estimate the corresponding TMRCA. The following clades were not selected for further analyses, Human H2N2, European Swine H1N1 and swine-origin pandemic H1N1, due to their small sample size or very recent origins. To ensure that the divergence times for the same clades were estimated in each polymerase tree, each clade was further examined for any incongruent isolates across the three polymerase trees. These isolates, which may represent reassortant viruses, were subsequently removed from the clade dataset. Thus, for each clade, the gene alignments comprised of the same taxa set. Further down-sampling was performed based on random selection procedure, so that not more than 100 isolates were included in each dataset, whilst maintaining the temporal variation of the original datasets.

The TMRCA estimates for each clade were estimated from the three polymerase alignments in a single BEAST analysis (Drummond and Rambaut, 2007). To improve the precision of the TMRCA estimate of the clade, the analysis was repeated by em-

Figure 5.1: Maximum likelihood trees for the polymerase genes of influenza A viruses. A sample 891 isolates from human, swine and avian isolates are included in these trees. Branches marked by * denote support ¿70%. All trees have been midpoint rooted for purpose of visualization.



ploying a hierarchical Bayesian prior (Suchard et al., 2003) on the tree root heights (which effectively pools the estimates from the three polymerase genes) to provide a joint estimate for the TMRCA. In the single BEAST analysis, for each partition (i.e. the polymerase genes: PB2, PB1 and PA) a separate codon-structured SDR06 substitutional model (Shapiro et al., 2006b), a relaxed uncorrelated lognormal clock (Drummond et al., 2006) and a GMRF coalescent prior (Minin et al., 2008) was applied. Either specific time information (day, month and year) where readily available or the year of isolation were used as sampling dates. For Human H1N1 (1918-) isolates that were collected after 1977, an approximate correction of 26 years was applied to account for a period of stasis in the molecular clock (Nakajima et al., 1978; Scholtissek et al., 1978; Wertheim, 2010). The MCMC chains were run at 100 million generations and were sampled regularly to yield a posterior distribution based on 10,000 tree estimates. Multiple MCMC chains were executed to evaluate convergence, which were subsequently combined to increase the number of posterior samples.

Influenza B and C viruses

Since influenza B and C viruses are not characterised by distinct antigenic subtypes like influenza A, and were mostly sampled from humans, a simpler procedure was carried out to collate the datasets. Viral isolates where all segments were sequenced, with dates of isolation, were downloaded from GenBank and aligned manually. To determine the TMRCA of each virus, the same BEAST settings applied previously to the influenza A viral clade were employed.

Long-term divergences

Dating the MRCA of influenza A, B and C viruses

For the older divergences between the influenza virus types, an amino acid substitutional model was used to estimate the evolutionary distances between the influenza viruses. As these models are more computationally intensive to run in a Bayesian phylogenetic framework, only a subset of isolates (ranging from 3 to 5) was selected for

each group. The two datasets, influenza A, B and C (ABC) and influenza A and B (AB), each contained 10 viral isolates. The amino acid alignments for each polymerase gene (PB2, PB1 and PA/P3) were manually assembled from the translated nucleotide sequences.

The TMRCA date estimates obtained for each group from the short-term divergence analyses were applied as informative priors for the corresponding nodes in the AB and ABC trees. The dates of divergence of AB and ABC viruses were jointly estimated from the polymerase gene alignments, with and without a hierarchical Bayesian prior (to improve the estimate) on the TMRCA of each influenza virus (A, B and C). The molecular dating analyses were performed in BEAST, using WAG amino acid substitution model (Whelan and Goldman, 2001) with gamma distributed rates and a relaxed uncorrelated lognormal clock (Drummond et al., 2006). In addition, since the demographic history of these viruses is unknown a priori, and not expected to be sufficiently informative for the molecular clock estimation, a diffuse GMRF coalescent prior (Minin et al., 2008) was employed. Multiple MCMC runs of 100 million generations were executed to increase sampling and to evaluate adequate mixing and chain convergence.

5.3 Results

Short-term Divergences

The TMRCA for the influenza A virus clades estimated in this study using the hierarchical Bayesian model are shown in Table 5.1. The dates of divergence of the human virus lineages closely agreed with estimates based on the polymerase genes inferred by Smith et al. (2009a). Interestingly, the divergence of the Human H1N1 and Classical Swine lineages was estimated as between 1911 and 1917, even though the TMRCA of the Human H1N1 lineage was estimated between 1888 and 1916 (Table 5.1). This lack of time resolution most likely reflected the absence of viral sequences isolated during the early emergence and divergence of these lineages. The oldest viral isolate included in

Table 5.1: Table of TMRCA estimates for the influenza A virus clades identified in Figure 5.1

Clade	Temporal range of samples	Joint TMRCA estimate
Influenza A (**All hosts**)	1918-2009	1900.48 (1913.13, 1881.34)
Human-Swine H1N1 split	1935-2009	1914.61 (1911.27, 1917.28)
Classical Swine	1935-2009	1932.83 (1930.13, 1935.30)
Human H1N1	1918-2008[a]	1903.81 (1888.51, 1916.11)
Human H3N2	1968-2008.175	1966.30 (1964.74, 1967.92)
Eurasian	1996-2009.123	1979.77 (1974.06, 1985.25)
North American	1982.614-2006	1941.66 (1918.46, 1969.05)

[a]Given the 26 years of stasis, this sampling range represents an evolutionary scale between 1918 and 1982.

this study was sampled from a human in 1918 (A/Brevig Mission/1918/H1N1). Its phylogenetic placement in different segment trees led Smith et al. (2009a) to the conclusion that it mostly likely represented a reassortant virus, deriving from a pre-1918 Human H1N1 lineage and new avian influenza A virus. Although only polymerase genes were analysed in this study, the isolate A/Brevig Mission/1918/H1N1 consistently clustered with the seasonal Human H1N1 clade in the maximum clade credibility (MCC) trees when no phylogenetic constraints were applied. The posterior support for the monophyletic grouping of Human H1N1 ranged from 0.47-0.70, indicating that in contrast to Smith et al. (2009a), this evidence of inconsistent clustering pattern of A/Brevig Mission/1918/H1N1 was not strong. Without additional viral sequence samples between 1918 and 1930, it is unlikely we can resolve the phylogenetic origin of the oldest influenza A virus isolate with any certainty. The date of divergence of the classical swine lineage overlapped with estimates from Dunham et al. (2009), although the estimates obtained in this study were more precise (smaller HPD intervals). The avian influenza A virus clades were estimated with surprisingly younger dates of divergence, with mean TMRCA of 1941 and 1980 for the North American and Eurasian clades respectively. Since the selection procedure for including isolates for each clade enforced monophyly across the polymerase gene trees, this may have biased the sampling range

to more recent dates (Chen and Holmes, 2010; Suzuki and Nei, 2002; Wertheim and Kosakovsky Pond, 2011). The earliest avian influenza sample in the full 891 data set was sampled in 1949, however it was not included in the analysis due to the inconsistent clustering between the North American and Eurasian avian clades. The divergence of all influenza A viruses was estimated between 1881 and 1913, implying a more recent MRCA in comparison to estimates obtained from antigenic genes (Chen and Holmes, 2006, 2010; Suzuki and Nei, 2002; Wertheim and Kosakovsky Pond, 2011). Although this may reflect the sampling bias as explained previously, it is interesting that the TMRCA of the avian influenza A viruses estimated by Chen and Holmes (2010), based on the polymerase genes, was found to be of similar range, between 1848 and 1935.

Short-term divergences of influenza A, B and C

Figure 5.2 illustrates the mean (joint) TMRCA age of the influenza A, B and C viruses, which were estimated independently with nucleotide substitutional models, and from the AB and ABC trees using the amino acid substitutional models. Apart from influenza A viruses, the divergence time estimates for the virus groups did not differ significantly between the independent analyses based on nucleotide sequences, compared to those estimated from the AB and ABC trees using amino acid sequences (Figure 5.2). For the date of divergence of influenza A viruses, there appeared to be an effect between the nucleotide and amino acid sequences based estimates, where the age of TMRCA(A) increased with the latter. For example, the divergence time based on nucleotide sequences estimated a mean TMRCA(A) of 1900.48, which increased to 1891.66 and 1876.62 with the AB and ABC tree respectively (Figure 5.2).

Long-term divergences of influenza viruses

The divergence times of influenza A and B (AB) and all influenza viruses (ABC) were estimated using amino acid substitutional models, which are expected to be comparatively robust to saturated sites when inferring long evolutionary distances. The dates of divergences for both AB and ABC were estimated with great uncertainty, with the

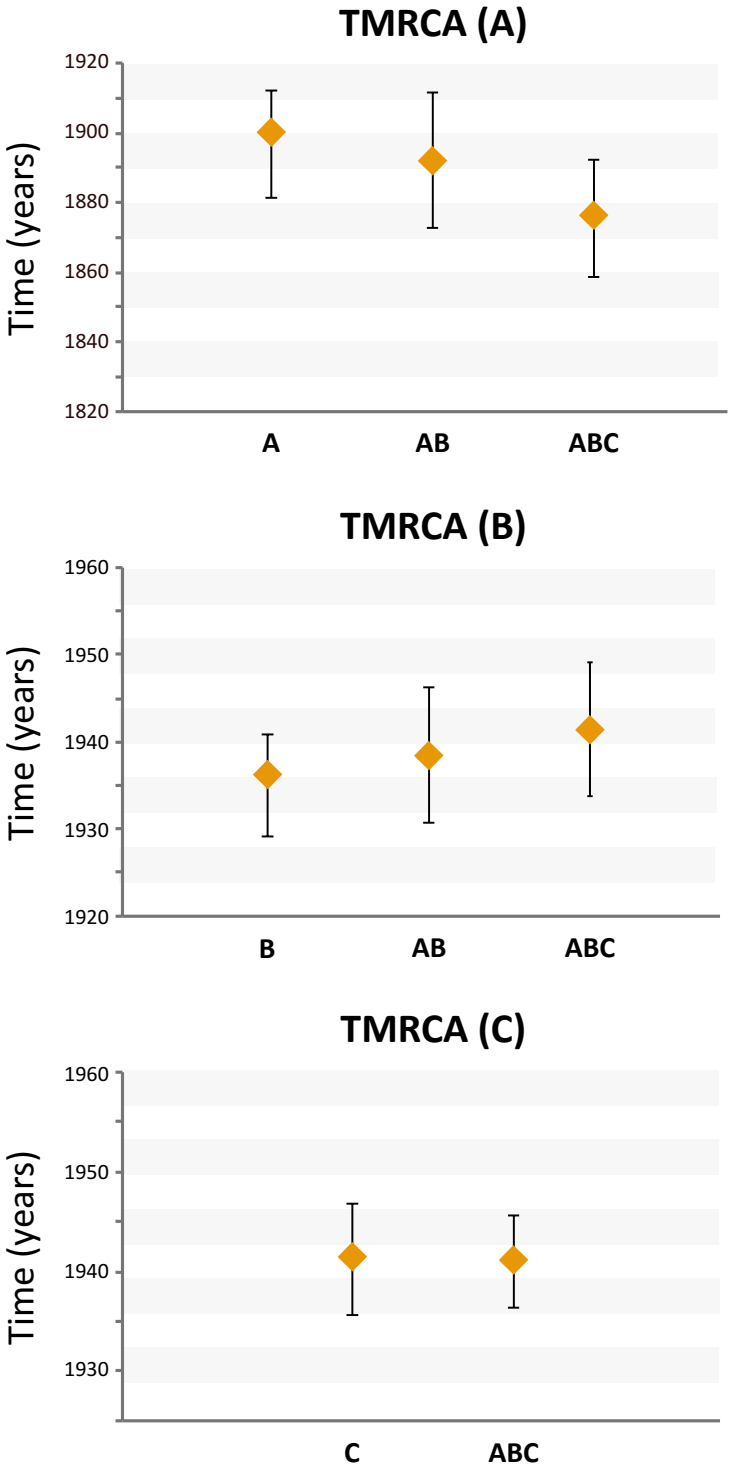


Figure 5.2: TMRCA estimates of influenza A, B and C. These have been inferred independently with nucleotide substitution models (first estimate in each plot), and from the AB and ABC trees using amino-acid substitution models. The error bars depict the 95% high posterior density interval of the estimates.

95% HPD intervals for the two respective nodes ranging from 103-3130 and 297-6335 years ago. Nevertheless, the upper bound for the ABC divergence time was estimated to be much older than upper bound of the AB divergence, indicating that there is sufficient power to differentiate the two events phylogenetically, but not temporally with any significant statistical certainty.

5.4 Discussion

The results indicate that although short-term divergences of influenza viruses can be accurately recovered with nucleotide sequences, the long-term divergence estimates are associated with large statistical uncertainty, even when amino acid substitutional models are employed. This discrepancy suggests that the amino acid substitution models are not adequate to model long-term evolutionary rates of fast-evolving RNA viruses. In contrast to short-term evolutionary rates, where the nucleotide substitutions can be realistically modelled by a continuous-time Markov process, to accurate estimate long-term evolutionary rates: additional consideration of selective pressures acting on the sequences is expected to be necessary. Over short evolutionary timescales, RNA viruses evolve relative quickly, with observed changes in their genomes comprising mostly neutral or nearly neutral substitutions (Pybus and Rambaut, 2009). However, the degree of sequence identity maintained between divergent viral lineages suggests that long-term evolution must be affected by strong selective constraints to ensure functionality of the encoding genome (Edwards et al., 2006; Holmes and Twiddy, 2003; Pybus et al., 2007b). Therefore, while we can safely assume that the substitution process is identically and independently distributed among sites in the short-term, it is unlikely to apply in the long-term. Structural constraints imposed either at the RNA or protein level, along with epistatic interactions suggests that the substitution process over long evolutionary timescales is expected to complex and non-independent among sites (Holmes, 2003a; Sanjuán et al., 2005). A recent study by Wertheim and Kosakovsky Pond (2011) has demonstrated that by incorporating the effects of purifying selection into codon substitutional models, dates of ancient viral divergences could be pushed further back in

time than current estimates based on nucleotide and amino acid substitution models.

The short-term divergences of influenza viruses closely agree with previous studies that have used polymerase genes for molecular clock dating (Chen and Holmes, 2006, 2010; Dunham et al., 2009; Smith et al., 2009b). In particular, the recent date of divergence of the influenza A viruses is interesting, given that estimates based on the antigenic genes give much older divergence times (Chen and Holmes, 2006, 2010; Suzuki and Nei, 2002). This pattern of low divergence among the internal segments, which includes the polymerase genes, and high divergence among the external (antigenic) segments, has been proposed by Chen and Holmes (2010) to be driven by a serial hitchiking effect of the viral genome when the antigenic genes periodically undergo selective sweeps. Since the same antigenic genes are unlikely to be involved in each selective sweep event due to herd immunity in the population, the genetic diversity of the associated internal segments will be disproportionately reduced (Chen and Holmes, 2010). This hypothesis also relies on frequent reassortment of the internal segments, which facilitates the dissemination of the alleles through the population (Chen and Holmes, 2010). However it is important to note that the TMRCA estimates obtained by (Chen and Holmes, 2010) for all gene segments were based on nucleotide substitutional models. Therefore, it will be interesting to test their observation with more realistic models of substitution that consider varying selective pressures, and see whether we can reconcile the dates of divergence of influenza A viruses across the genome (Webster et al., 1992).

The divergence of influenza viruses presents a challenging problem for current evolutionary analyses. An alternative approach developed by O'Brien et al. (2008) has estimated the TMRCA of influenza A and B viruses existing just only 100 years ago. While this study has taken steps to investigate dating methods that do not rely on the molecular clock, it is difficult to imagine that the three viral lineages of influenza (A, B and C) have occurred over relatively short evolutionary periods given that they differ phenotypically and circulate among different hosts. Furthermore, the lack of reassortment among the lineages suggests that these viruses have diverged over longer

evolutionary periods.

The hierarchical analytical framework introduced here has allowed to evaluate the suitability of substitutional models to date ancient viral divergences among fast-evolving RNA viruses. The imprecise estimates obtained for the ancient divergence of influenza viruses indicate that current models of evolutionary change are inadequate to infer long evolutionary distances. This is most likely due to these models failing to account the long-term evolutionary dynamics of acute-infecting RNA viruses, where natural selection is expected to be a major determinant. This study highlights that to investigate ancient evolutionary timescales of RNA viruses, we need better models of sequence evolution that consider the factors shaping their long-term evolution.

6

DENV-1 Transmission in Southeast Asia

The manuscript form of this chapter has been published in ***PLoS Pathogens 7, e1002064*** as, "Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission", (Raghwani et al., 2011).

Viral Sequencing and data collection was carried out by V. T. Hang, T. T. Hien, J. Farrar, B. Wills, N. J. Lennon, B. W. Birren, M. R. Henn, and C. P. Simmons.

I analyzed the data and wrote majority of the manuscript.

C. P. Simmons, E. C. Holmes, and A. Rambaut provided editorial assistance.

6.1 Introduction

Dengue is a global emerging disease, with the number of cases and countries affected increasing annually (Initative, 2009). The co-circulation of multiple serotypes in endemic areas has also seen progressively severe outbreaks, with rising incidence of DHF/DSS cases among the infected population (Guzmán and Kourí, 2002). To control dengue prevalence, a better understanding of how it spreads in endemic populations is required.

The incidence of dengue varies both spatially and temporally, and is driven by a combination of factors involving the human host, the mosquito vector and the virus(Endy et al., 2004; Holmes and Twiddy, 2003; Rabaa et al., 2010; Schreiber et al., 2009). The proportion of partially immune and susceptible individuals is important in shaping the transmission dynamics of dengue (Anders et al., 2011; Nagao and Koelle, 2008), especially in determining the cycling pattern of serotype circulation in hyperendemic areas (Adams et al., 2006). The vector population dynamics are also a key determinant of dengue disease patterns; the population size and density are affected by climate-related factors such as levels of rainfall and temperature (Scott et al., 2000; Scott and Morrison, 2003). Lastly, differences in the virus traits, such as virulence and fitness, are also likely to contribute to patterns of disease transmission (Bennett et al., 2006; Cologna et al., 2005; Hang et al., 2010; Holmes and Twiddy, 2003; Libraty et al., 2002; McElroy et al., 2011; Messer et al., 2003; Myat Thu et al., 2005; Rico-Hesse et al., 1997).

At the local level, dengue transmission is typically associated with focal spread in both urban and rural environments (Jarman et al., 2008; Mammen et al., 2008; Morrison et al., 2010; Schreiber et al., 2009). This implies that either the human host or the mosquito vector do not move sufficiently far, such that strong spatial structure in the dengue viral diversity is often observed, even in small regions (Jarman et al., 2008; Mammen et al., 2008; Morrison et al., 2010; Schreiber et al., 2009). In addition, dengue transmission is associated with frequent viral migration, both at local and larger geographic scales (Foster et al., 2003; Schreiber et al., 2009). Since these findings have come from spatially and temporally confined studies, it is not clear what the exact roles

of human and mosquito movement are in more localised dengue transmission.

To investigate these questions, this chapter focuses on DENV-1 transmission in South East Asia in the context of the recent DENV-1 epidemics in southern Viet Nam. Southeast Asia is a very appropriate setting for this study as it has one of the highest dengue prevalences in the world, and provides an opportunity to examine dengue transmission at different geographic scales. Furthermore, utilising an expansive dataset that has been sampled both spatially (rural and urban locations) and temporally (between 2003 and 2008) from southern Viet Nam, the fine-scale transmission dynamics of dengue are inferred using the recently developed Bayesian phylogeographic models (Lemey et al., 2010, 2009) implemented in BEAST (Drummond and Rambaut, 2007). This unique framework and comprehensive analysis has allowed us to gain detailed insight into both the ecological and population dynamics of dengue at local, regional and international scales.

6.2 Methods

Information about the patients and genome sequencing of the DENV-1 isolates sampled during 2003-2008 from southern Viet Nam can be found in Appendix C.

Phylogeographic analyses of DENV-1 in Southeast Asia and Viet Nam

A data set of DENV-1 sequences was collated to include isolates from countries in Southeast Asia that are likely linked to Viet Nam via migration. An alignment of the envelope (E) gene (1485 nt) was assembled for the Southeast Asian and Vietnamese isolates (n=134 and 751, respectively) to include the broadest range of locations. An initial neighbor-joining tree was constructed in PAUP* (Swofford, 2003), using a HKY85 nucleotide substitution model with gamma-distributed rates. This allowed us to make an initial identification of the major clades of DENV-1 in Viet Nam. These Vietnamese isolates were then subsampled (n=101) to explore their phylogeography in context of the South East Asian isolates. Isolation dates for the South East Asia data set were obtained from GenBank annotations and via personal communication. Where specific

dates were not available in terms of day and month, a mid-point of the year of isolation was used. The spatial dynamics of DENV-1 in Southeast Asia were investigated with a discrete diffusion model (Lemey et al., 2009) using Bayesian Monte Carlo Markov Chain (MCMC) method implemented in BEAST (Drummond and Rambaut, 2007). The phylogeography analysis was executed with a codon-structured SDR06 substitution model (Shapiro et al., 2006a), a relaxed uncorrelated lognormal clock (Drummond et al., 2006) and a Gaussian Markov Random Field (GMRF) coalescent prior (Minin et al., 2008) over the unknown phylogeny. The discrete diffusion model uses the country of isolation of the sampled sequences to reconstruct the ancestral location states of the internal nodes from the posterior time-scaled tree distribution. The MCMC chains were run for 50 million generations, sampling every 5000th state, and were executed multiple times to ensure adequate mixing and stationarity had been achieved.

Viral transmission between Ho Chi Minh City and Dong Thap province

Major clades of Vietnamese DENV-1 identified from the broad-scale South East Asian analysis were selected for further study to examine the spatial and temporal variation in Viet Nam. In clades with appreciable numbers of sequences from Dong Thap and HCMC, isolates were analyzed independently to gauge the regional variation in viral transmission patterns. For the fine-scale analysis, a continuous diffusion model based on a lognormal relaxed random walk (Lemey et al., 2010) was employed to model the DENV-1 spatial dynamics in Viet Nam. For each isolate, the specific sample dates and location information with respect to the longitude and latitude of the patients household were used. Isolates that were identical in sample dates and location information were down-sampled to reduce the potentially biasing effect of over-sampling of epidemiologically-linked cases. The MCMC runs were run and evaluated as previously described, and the chain lengths ranged from 50 to 100 million generations, and sampled regularly to yield 10,000 trees from the posterior distribution.

The viral dispersion rates (km/yr) for each data set were calculated across the tree (i.e. total straight-line distance travelled divided by the total time) and biannually to

consider the spatial heterogeneity in a time-scaled framework. Plots of relative genetic diversity over time were reconstructed using the GMRF coalescent prior to reveal the association between the genetic diversity of each group in terms of their evolutionary history (Minin et al., 2008). Further discrete phylogeography analyses were performed with the robust counting method (Minin and Suchard, 2008; O'Brien et al., 2009), to determine the extent of viral migration between Dong Thap and HCMC and whether this varied when the lineage originated in a rural or urban area. In this case, the discrete states were represented by either the isolate being sampled from HCMC, Dong-Thap or neither (non-Dong Thap or HCMC).

Viral Migration within Ho Chi Minh City

For the limiting case of a freely mixing (non-spatially structured) epidemic in HCMC, dispersion rates were estimated whilst randomizing the tip locations during the tree proposal in the MCMC, whilst co-estimating the rates for each independent lineage and the joint DENV-1 diffusion rate. To determine the viral transmission network within HCMC, a non-reversible discrete phylogeography model was applied to all the HCMC isolates, using the district of isolation for the discrete states. The analysis was performed and evaluated as described above with the addition of Bayesian Stochastic Search Variable selection (BSSVS) was implemented to identify significant transition rates between locations (Lemey et al., 2009). The transition rates supported by a Bayes factor of at least 3 were examined further by looking at the number of in-degree and out-degree per district. The number of connections was normalized by the number of samples from the source location in order to reduce the bias from under-represented locations in our data set.

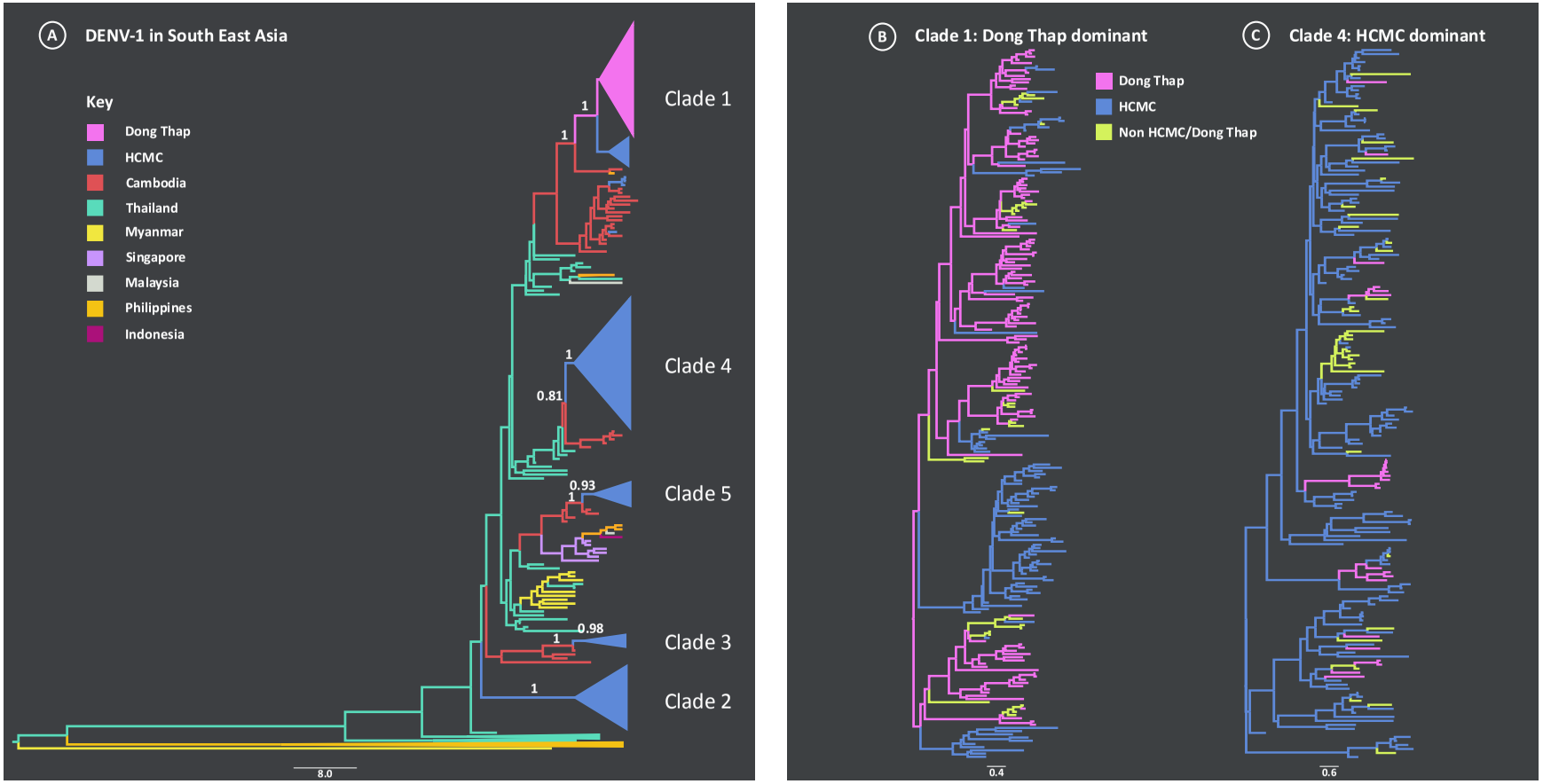


Figure 6.1: . Maximum Clade Credibility (MCC) trees from the discrete phylogeography analysis of DENV-1: A) South East Asia tree reconstructed from the E gene, where the branches are colored by location of viral samples. The number above the branch indicates the posterior probability support for the Vietnamese clades and with nearest sister clades; B) An in-depth look at Vietnamese clades 1 and 4 which were found to be Dong Thap dominant and HCMC dominant, respectively. The branches are colored by sampling location (Dong Thap, HCMC or Non-HCMC and Dong Thap).

Table 6.1: Rate of nucleotide substitution of DENV-1 for each clade in Viet Nam, with the 95% HPD intervals, and the inferred time of the most recent common ancestor (TMRCA).

Clade	Mean TMRCA	Rate of Evolution (10^{-3} subs/site/year)
1	2005.22	0.878 (0.776, 0.980)
2	2002.35	0.966 (0.864, 1.071)
3	2003.69	0.949 (0.622, 1.279)
4	2002.94	1.031 (0.912, 1.155)
5	2003.02	0.882 (0.656, 1.086)

6.3 Results

Phylogeography of DENV-1 in South East Asia

We determined the consensus DENV-1 genome sequence (minimum sequence from nt 70-10,400) in acute plasma samples collected from 751 hospitalized patients in urban Ho Chi Minh City (HCMC) (n=575 sampled between 2003-2008) and rural Dong Thap Province in the Mekong Delta region (n=176 sampled between 2006-2007). The majority of viruses were sampled from 2006 to 2008 during which DENV-1 was the most prevalent serotype in circulation (see Figure S1 in Raghwani et al. (2011)).

To determine the evolutionary relationships of DENV-1 in Viet Nam in the context of surrounding countries we analyzed the envelope (E) gene sequences from these locations (Figure 6.1 A). The 751 DENV-1 sequences sampled from Viet Nam fell into one of five clades within the broader Genotype I cluster of viruses (Zhang et al., 2005). Four of the five clades consistently clustered within the diversity of Cambodian viruses with good support (posterior probability ranging from 0.81 to 1.0). This phylogeographic evidence, coupled with Cambodia and Viet Nams shared border, is compatible with Cambodia acting as the major source of Vietnamese DENV-1. A caveat to this is the lack of contemporaneous DENV-1 sequences from nearby Thailand, which has previously been shown to harbor substantial DENV diversity and importation into Viet Nam (Hang et al., 2010). Clearly, wider sampling in both time and space is needed to test this hypothesis.

The majority of the clades largely comprised of viruses from HCMC, with the exception of Clade 1, which was found to be Dong Thap dominant. The timing of these inferred introductions can be gauged from the age of the most recent common ancestor (TMRCA) of each clade (Table 6.1). The period in which these different viral clades emerged in southern Viet Nam ranges from late 2001 to mid-2005. Apart from Clade 1, which was found to be the most recent introduction, the mean ages of clades 2-5 did not differ significantly, suggesting that different viral lineages were imported over short or similar time-scales, and then co-circulated. These clades were chosen for more detailed phylogeographic analysis. Finally, genome-wide rates of nucleotide substitution at 1×10^{-3} nucleotide substitutions per site, per year (Table 6.1) were the same among clades and highly consistent with those previously determined for DENV (Lanciotti et al., 1997; Twiddy et al., 2003).

Viral Migration between Dong Thap and HCMC

For the clades identified as being within Viet Nam, a discrete spatial model (Lemey et al., 2009) was employed to reveal the migration between the discrete sampling locations. The results are shown in Figure 6.1 B, in which branches are colored by the most probable state location. In four of the five clades HCMC was the most likely source of virus, with viruses exported to the rural area of Dong Thap. The non-HCMC isolates in these clades were interspersed among the HCMC sampled isolates, which strongly suggests that the DENV-1 epidemics in southern Viet Nam studied here mainly emerged first in HCMC. The exception was clade 1, which was dominated by Dong Thap viruses and where Dong Thap was inferred to be the most likely place of origin. Moreover, the HCMC isolates in clade 1 did not form a monophyletic group, supporting the view that clade 1 viruses were imported into HCMC on multiple occasions from Dong Thap.

To determine whether the viral migration rates varied between urban and rural epidemics, we compared spatial dynamics in clades 1 and 4 (Table 6.2). When focusing on the number of transitions from the inferred source location, a symmetrical pattern

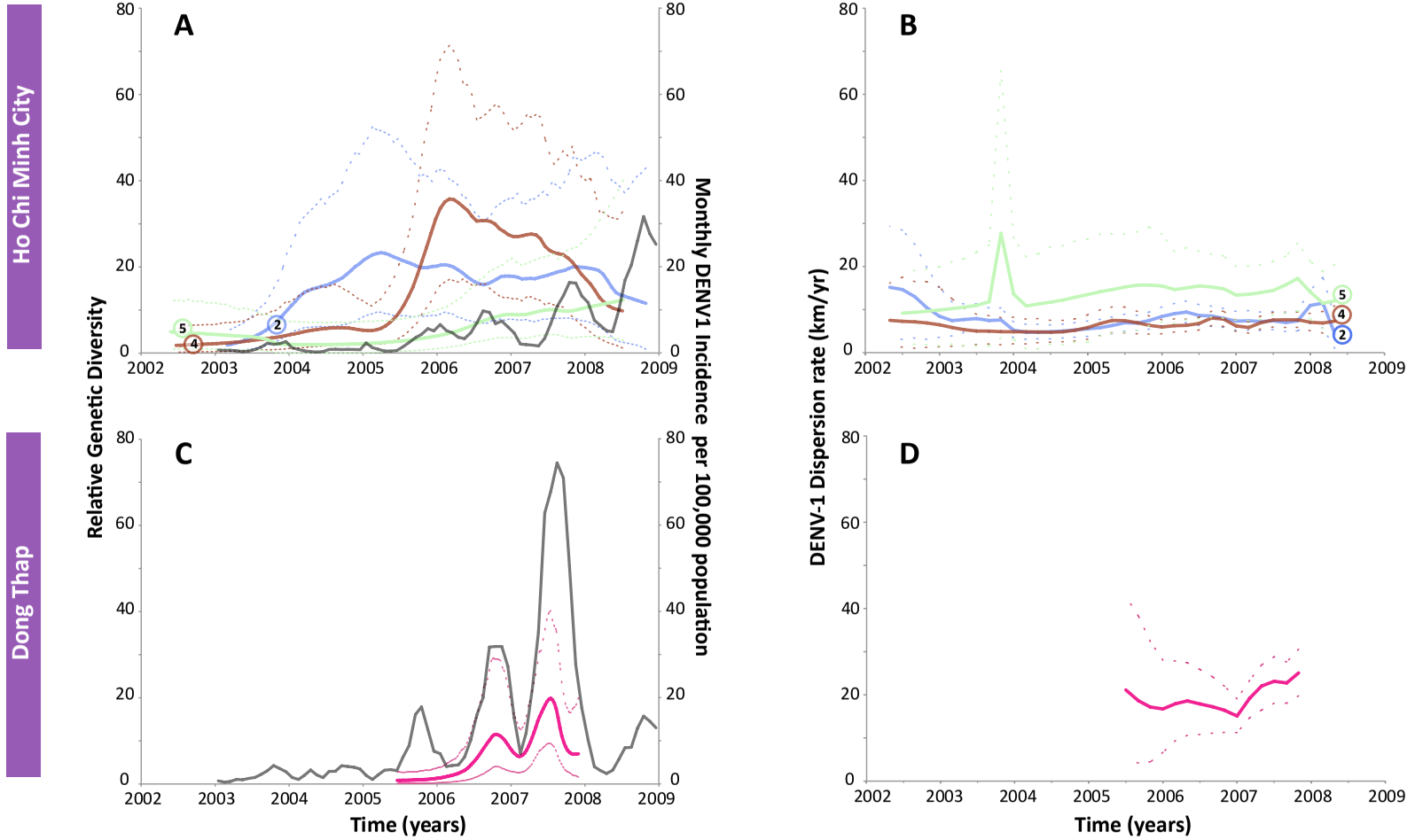


Figure 6.2: The genetic diversity, dengue incidence rate, and dispersion rates of DENV-1 in HCMC and Dong Thap. The top row shows the results of three dominant lineages in HCMC (clades 2, 4 and 5), while the bottom row shows the results of Clade 1 in Dong Thap. Clade 1 is not included in HCMC since it was identified as having multiple origins from Dong Thap. A) Relative genetic diversity of the main clades in HCMC (clades 2, 4, and 5 denoted by blue, red, and green respectively), superimposed with the incidence rate of DENV-1 in HCMC (dark grey). B) Dispersion rates of the three lineages estimated in 6-month intervals from 2003-2009. C) and D) show the results of Clade 1 in Dong Thap, relative genetic diversity (pink) and incidence rate (dark grey) and dispersion rate, respectively.

Table 6.2: Migration between HCMC and Dong Thap in Clade 1 and Clade 4, the number of transitions between each state along a branch in the tree using the robust counting method.

Clade 1	HCMC	DT	Non DT/HCMC
HCMC	-	1.20	5.28
DT	12.36	-	11.28
Non DT/HCMC	4.73	4.50	-

Clade 4	HCMC	DT	Non DT/HCMC
HCMC	-	12.60	21.93
DT	2.55	-	2.97
Non DT/HCMC	4.26	1.83	-

was observed between the two clades. For instance, the transmission rate between HCMC and Dong Thap was higher in the HCMC dominant clade 4, while for the reverse direction (Dong Thap to HCMC) it was greater in Dong Thap dominant clade 1. Hence, once a virus becomes established in a location, rural or urban, the rate of viral exportation was found to be greater than the rate of viral importation.

DENV-1 dispersion in urban and rural locales

The geographical coordinates of the patients residential address in HCMC (n=381) or Dong Thap Province (n=175) was known for 556 cases and this information was employed to reconstruct the fine-scaled dispersion of the individual viral lineages within the sampling areas using a continuous spatial diffusion model with non-homogenous dispersion rates (Lemey et al., 2010). The average viral dispersion rates (km/year) were calculated for each clade, and separately for HCMC and/or Dong Thap data subsets, as if the epidemic in these regions derived from a single introduction (Table 6.2). We define virus dispersion rate as a measure of how quickly a virus lineage spreads geographically, given the inferred root location and final sampling locations. Even though we only have one estimate of the average dispersion rate of DENV-1 in Dong Thap, a clear disparity was observed when compared to the rates from HCMC lineages (Table 6.3). Specifically, the viral lineages from clade 1 in Dong Thap spread

Table 6.3: Rate of virus dispersion of DENV-1 in the different localities in Viet Nam.

	Clade	Mean Dispersion Rate (km/yr)	Mean TMRCA
1	Dong Thap	20.61 (16.72 , 24.91)	2005.42
	HCMC	11.59 (9.10, 13.96)	2005.11
	All	28.68 (25.78 , 32.12)	2005.22
2	HCMC	7.32 (6.11 , 8.69)	2002.30
	All	12.44 (10.59 , 14.56)	2002.35
3	All	38.15 (24.31 , 52.53)	2003.69
4	Dong Thap	22.18 (10.39 , 34.56)	2003.20
	HCMC	6.78 (5.66 , 7.98)	2002.43
	All	18.55 (15.88 , 21.48)	2002.94
5	HCMC	14.37 (9.43, 20.10)	2002.33
	All	23.06 (15.94 , 29.95)	2003.02

approximately 2-3 times faster than any lineage from HCMC. This is indicative of a fundamental difference in the epidemiological dynamics of DENV-1 in the two areas.

A further dissection of the dispersion rates through time in HCMC (clades 2, 4 and 5) and Dong Thap (clade 1) revealed interesting patterns in the rate of viral spread in the two locations. In HCMC (Fig 6.2 A and B), the monthly incidence of DENV-1 showed a similar trend as in Dong Thap, with corresponding regular fluctuations and an increasing overall trend. However, there was no clear association between genetic diversity, incidence, and dispersion rate observed in the urban environment demonstrated by the roughly horizontal relationship in Figure 6.2 B and the overlapping 95% HPD (highest posterior density) intervals. Hence, although the DENV-1 clades were introduced independently into HCMC, they have spread at similar and effectively constant rates. For Dong Thap, clade 1 was the only one clearly derived from a distinct single importation and of a sufficient size for analysis. The dispersion rate of DENV-1 appeared to be associated with the fluctuations in genetic diversity and monthly incidence in Dong Thap (Figure 6.2 C and D). The two peaks in relative genetic diversity

Figure 6.3: The dispersion of the main clades in HCMC from 2003-2008 estimated from the continuous diffusion phylogeography process. The different clades are color coded (clade 2 = blue, clade 4 = red, and clade 5 = green). All three clades appear to emerge from similar parts of HCMC and continue to co-exist in same geographic space at the same time.

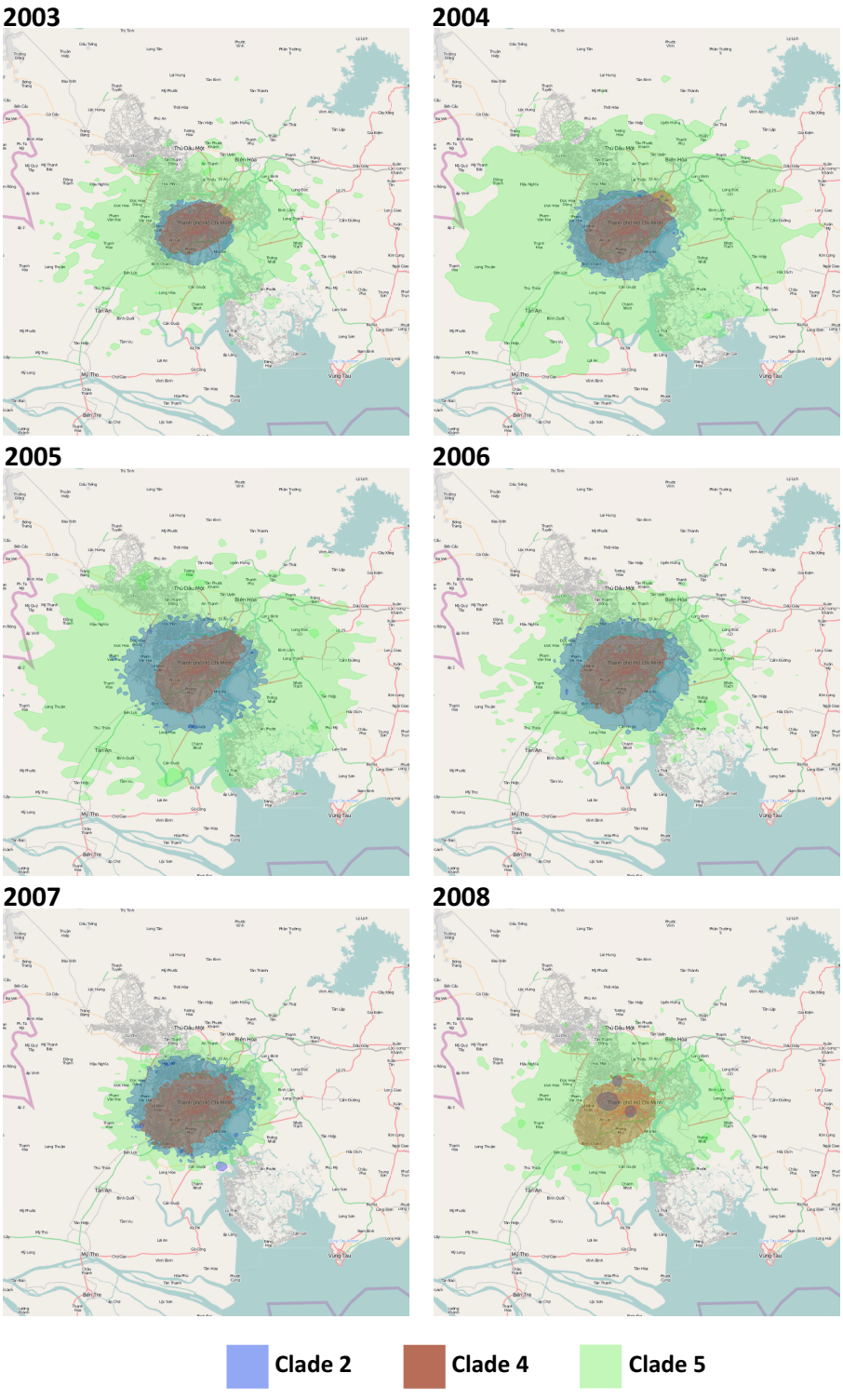


Table 6.4: Results from randomizing locations at the tips to test the upper limits of the dispersion rates of DENV-1 in HCMC, Viet Nam. The numbers inside the parentheses shows the 95% HPD intervals

	Empirical	**Randomization**
Clade	Mean Dispersion Rate (km/yr)	Mean Dispersion Rate (km/yr)
HCMC2	7.32 (6.11, 8.69)	21.10 (4.84, 155.21)
HCMC4	6.78 (5.66, 7.98)	14.32 (4.00, 85.69)
HCMC5	14.37 (9.43, 20.10)	39.68 (7.84, 216.01)
joint	7.49 (6.06, 9.31)	20.03 (5.45, 92.38)

of clade 1 in Dong Thap coincide with the two major peaks in the monthly incidence, indicating that DENV-1 epidemic in Dong Thap is largely driven by this lineage.

To investigate whether these dispersion rate estimates in HCMC were simply a reflection of the geographic constraint of our samples, they were re-estimated by randomizing the tip location for each clade (Table 6.4). The results indicate the maximum possible dispersion rate given the sampled locations, which were found to be 2-3 times greater than the empirical estimates, with wide HPD intervals (Table 6.4). The spatial reconstruction of the viral spread at different stages of the epidemics shows that these viral lineages have co-circulated in the same place at the same time (Figure 6.3). This observation is of fundamental importance as it suggests that the number of susceptible hosts to DENV-1 has not been saturated in HCMC, and could potentially support additional DENV-1 lineages in this area.

Population density and transmission routes

To determine whether transmission routes within HCMC varied according to population density, a non-reversible discrete phylogeography model was applied to district level data. Importantly, the more densely populated inner city districts (above 30,000 people per km^2) were found to contribute significantly to DENV-1 transmission compared to the suburban districts (Figure 6.4). Moreover, the most densely populated region, District 5, had the highest number of connections, providing compelling evidence that this area might be a major hub in the city.

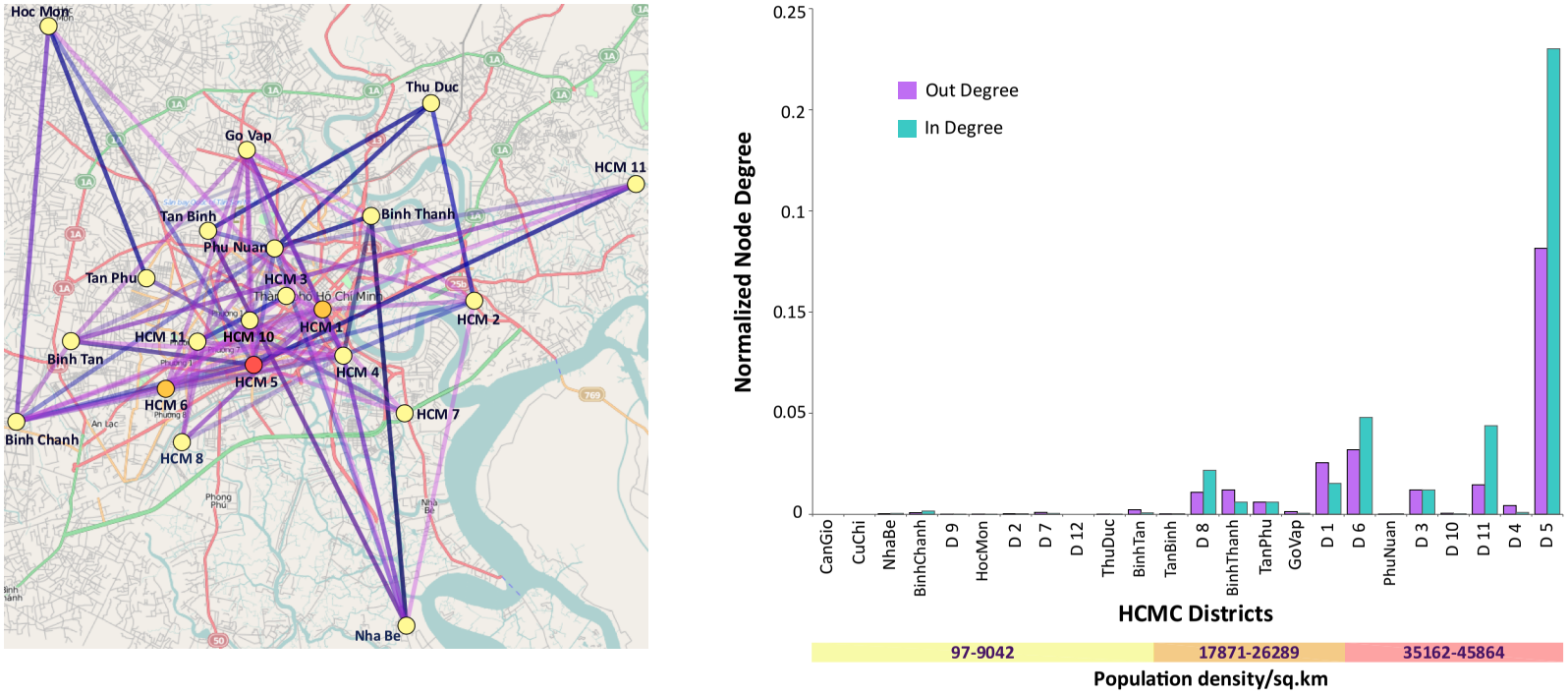


Figure 6.4: Results from a non-reversible discrete phylogeography analysis of HCMC clades at the district level (21 states). The significant connections at Bayes factor 3 are shown on the map of HCMC. The districts are colored by the number of normalized out-degree connections, where yellow, orange and red indicate low, intermediate, and high, respectively. The number of significant connections to and from a location (district) is represented by the bar chart in terms of in degree and out-degree respectively. The districts along the x-axis are ordered by increasing population density.

6.4 Discussion

Endemic dengue in Southeast Asia is largely characterised by both restricted viral movement on a local scale with periodic importations from nearby countries. This results in the dengue viral diversity showing great spatial structure at the country level, indicating that there is a far high level of gene flow within countries than between countries. Further inspection of regional and local DENV-1 transmission patterns in southern Viet Nam suggest that host population density is major factor in driving the spatiotemporal dynamics of endemic dengue.

The introduction of at least five different DENV-1 lineages in southern Viet Nam indicates that human movement is important in shaping the extant viral diversity of endemic areas. While this study only investigated three DENV-1 clades in detail, the observation of multiple lineages circulating in the same regions and time, demonstrates that there are sufficient supply of susceptible hosts to maintain concurrent transmission. Furthermore, this pattern varies between urban and rural locations, which most likely reflect the differences in population densities. While the sequence data is biased towards the urban HCMC, the observation that the single DENV-1 lineage examined in rural Dong Thap recapitulated the epidemiological cases, indicates that rural population are unlikely to support as many viral lineages as urban locations.

Previous studies that investigated dengue in restricted regions, either small towns or villages, have reported a very focal transmission pattern, such that spatial structure in the dengue genetic diversity is observed in short periods of time. However, it has been unclear how quickly dengue viruses move over geographic distances, and what this reveals about the host and mosquito vector. Due to the detailed spatiotemporal sampling of our sequence data, the viral movement of DENV-1 lineages was measured directly in both Dong Thap and HCMC. The surprisingly low dispersion rates estimated in this study, especially in HCMC, indicates that the local transmission is attributed to mosquito rather than human movement. This is particularly striking since DENV-1 re-emerged in the area very quickly and rapidly disseminated across the population in

the early periods.

The high population density of urban HCMC (3024 person/km^2) and the restricted spatial movement of dengue indicate that mosquitoes do not need to move very far to infect new hosts. The elevated dispersion rates observed in rural Dong Thap, which has much lower population density (495 persons/km^2), is compatible with this hypothesis since mosquitoes would need to travel further to continues its transmission in sparsely populated areas. Although, since only one estimate for DENV-1 viral dispersion was obtained from the rural location, further investigation would be necessary to confirm whether urban and rural transmission consistently varies in this manner. Additionally, it is striking that the multiple lineages co-circulating in HCMC, have spread at effective the same rates through the population. The associated broadly similar genetic diversity and plasma viraemia levels of the different DENV-1 lineages, indicates that there is little or not fitness differences. Given these observations, it seems that the high population density of HCMC is likely to reduce any ecological competition between the strains due to the extremely large supply of susceptible hosts.

The population density of the locations also affects viral migration of endemic dengue within southern Viet Nam. For example, DENV-1 transmission tends to occur over a gradient of population density such that DENV-1 viruses tend to be exported from urban areas to rural areas. This gravity model of viral transmission is in line with host population density being a major determinant of how quickly the virus spreads on a geographical scale. A similar pattern has been previously observed with DENV-2 transmission in southern Viet Nam, where the urban population was found to have a central role in disseminating the virus throughout the region (Rabaa et al., 2010). Within HCMC, the highest viral movement is associated with the most densely populated region, District 5, indicating that areas of high population density drive focal transmission of dengue. It is also demonstrates that the low viral dispersion rates in HCMC does not mean that DENV-1 is moving slowly at the spatial scale, but that the viral movement is a function of the numbers of susceptible hosts, so in areas of high population density the mosquito does not need to move very far to find new hosts.

In spite of the high dengue prevalence in southern Viet Nam, spatial movement of the DENV-1 is greatly restricted and associated with focal viral transmission. These observations can be explained by the short travel distances of the mosquito vector, *A. aegypti*, which moves on average 100m over its life span of a few weeks (Harrington et al., 2005; Reiter et al., 1995; Russell et al., 2005), such that most DEN transmission at the local scale is mediated by mosquito dispersal rather than human movement. Moreover, this pattern of mosquito-driven transmission of dengue is largely dependent on the high density of susceptible hosts in endemic areas. A similarly limited movement of dengue has been reported by recent studies that focused on smaller geographic areas, which demonstrates the restricted spatial range of mosquito vectors, and further corroborated the highly focal pattern of DENV transmission observed in HCMC (Balmaseda et al., 2010; Schreiber et al., 2009).

The spatiotemporal patterns of the DENV-1 lineages in HCMC indicates that the full geographic range of DENV-1 lienages is established early on in the epidemic, which is compatible with the viruses disseminating rapidly through the population upon their introduction. A similar observation was found in Iquitos, Peru after the introduction of a new dengue serotype, where early-confirmed cases were scattered throughout the city, suggesting a rapid establishment of the virus when entering a completely nave population (Morrison et al., 2010). This observation gives added weight to our conclusion that the dispersion rates of DENV-1 in southern Viet Nam are a function of the availability of susceptible hosts.

While in this study, the susceptible host population did not appear to be saturated by multiple DENV-1 introductions, it will be interesting see when this balance changes such that the invasion of another dengue serotype is favoured in the population. Given these findings, it is clear the urban locations and their associated high population densities are major drivers of dengue transmission in endemic areas both at the local and regional levels. In terms of disease control, these results indicate that mosquito populations should preferentially target urban areas, in particular regions with the highest population densities, over rural areas.

7

Conclusions

By concentrating on fast-evolving RNA viruses with different genomic structures, transmission modes and ecology, this thesis has aimed to elucidate the patterns and processes that shape the complex dynamics of emerging viral diseases. A secondary aim has been to address some of the phylogenetic challenges that are presented by evolutionary analyses of emerging RNA viruses. In particular, I have looked at homologous genetic exchange and ancient divergence of human RNA virus pathogens. Therefore, while I have mostly focused on empirical analyses on viral sequence data, I hope to have also introduced improved phylogenetic-based techniques to examine fast-evolving RNA viruses.

Seasonal influenza is associated with frequent viral migration, both at the local and global scales. This observation has also coincided with detection of reassortant viruses, suggesting that the process of segmental exchange could be an important determinant in shaping seasonal viral diversity (Holmes et al., 2005; Nelson et al., 2007, 2008b,a). Furthermore, it may also mediate selective sweeps by bringing together favourable combinations of antigenic genes or other gene segments (Holmes et al., 2005; Rambaut et al., 2008). To understand this phenomenon better in influenza A viruses at the epidemic level, chapter 2 presents a method to quantify reassortment based on phylogenies estimated from different parts of the viral genome. The algorithm efficiently extracts the common evolutionary backbone between time-scaled tree distributions. This allows to assess the relative contribution of reassortment and phylogenetic uncertainty associated with the data in a statistical framework. Additionally, by focusing on shared evolutionary history of the viral genome, this method can also be applied to examine fine-scale evolutionary processes such as epistatic interactions and convergent evolution.

A related approach is employed in in chapter 4, which examines the origin of the unique HCV CRF strain, 2k/1b. Phylogenies constructed from different genomic regions of recombinant viruses describe independent evolutionary histories. However, in the case of CRFs, the node representing the most recent common ancestor (MRCA) of the clade is shared among the independent phylogenies. This means that the MRCA age of CRF01_1b2k can be jointly estimated from different gene sequences. By ex-

ploiting the data fully, this approach yields molecular clock estimates that are more precise. The example of CRF01_1b2k illustrates that rare recombination events in viral evolution, as observed with HCV, can be greatly informative about the underlying epidemiological processes by revealing that specific strains must have co-circulating. This is in stark contrast to reassortment in influenza A viruses and recombination in HIV, where these processes are prevalent and thus are associated with more complex evolutionary histories.

Chapter 3 demonstrates some of the inherent difficulties associated with investigations studying novel viral emergences, in this case SARS-CoV, when we have limited surveillance. The currently sampled viruses from the animal markets and bats are unlikely to represent the direct ancestors of the human SARS-CoV. However, the restricted temporal sampling of palm civets during the epidemic, and the clear undersampling of the ancestral viral reservoir in bats, means that we cannot conclusively eliminate their involvement in the emergence of the SARS outbreak. Additionally, this study highlights the problem of prior interactions in Bayesian phylogenetic investigations, which can severely bias evolutionary estimates and lead to incorrect inferences about the data. While the Bayesian framework has become a powerful tool to study fast-evolving RNA viruses, future investigations should be aware of the potential consequences of using restricted or informative priors.

The phylodynamic approach has been especially successful in examining the short-term evolution of emerging RNA viruses. This is largely due to their high evolutionary rates, a product of the underlying mutation rates and large population size, which result in observable change among viral sequences sampled over short time periods. Chapter 6 addresses whether we can apply this framework to investigate longer evolutionary timescales of fast-evolving RNA viruses in humans. In particular, I looked at dating the divergence of influenza viruses, A, B and C in humans. The divergence times estimated for the two ancient events (A, B and A, B and C) with amino-acid substitution models were indistinguishable, due to the large uncertainty associated with the estimates. The inadequacy of the amino-acid substitution model to infer ancient

viral divergences, which most likely reflects the long-term evolutionary constraints on the viral genome, indicates that more sophisticated models need to be considered when dating long evolutionary timescales. This includes models that can incorporate complex evolutionary processes, such as purifying selection, compensatory evolution and epistatic interactions.

In contrast to chapter 3, the findings presented in chapter 7 demonstrate the power of the phylogenetic method when viral sequence data has been appropriately sampled both temporally and spatially. Using an expansive dataset set collected from southern Viet Nam, a dynamic picture of how dengue circulates in endemic areas is revealed over different geographic scales. In particular, we find that host population density, which reflects the number of susceptibles in the population, is a major determinant of dengue viral transmission at both local and regional levels. Multiple viral lineages co-circulate within the urban environment, Ho Chi Minh City (HCMC), spreading at similar rates through overlapping geographic areas, suggesting that they are of equivalent fitness. Additionally, we observe that viruses generally disperse over a gradient of population density in both HCMC and between urban HCMC and rural Dong Thap, such that regions of high population density export viruses to regions of lower population density. Surprisingly, we also find that endemic dengue viruses are associated with low viral dispersion rates, especially in HCMC where the virus moves less than 20km/yr. These low rates are compatible with mosquito-mediated dispersal of dengue in local populations, which are characterised by large number of susceptibles. As a result, the mosquito vector (and the virus) in endemic areas does need to travel far to infect new hosts. Together these observations suggest that disease control measures where dengue is endemic should preferentially target regions of high population density.

Bibliography

Adams, B., Holmes, E. C., Zhang, C., Mammen, Jr, M. P., Nimmannitya, S., Kalayanarooj, S., Boots, M., Sep 2006. Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in bangkok. Proc Natl Acad Sci U S A 103 (38), 14234–9.

Aitken, C. K., Lewis, J., Tracy, S. L., Spelman, T., Bowden, D. S., Bharadwaj, M., Drummer, H., Hellard, M., Dec 2008. High incidence of hepatitis c virus reinfection in a cohort of injecting drug users. Hepatology 48 (6), 1746–52.

Alfaro, M. E., Holder, M. T., 2006. The posterior and the prior in bayesian phylogenetics. Annu Rev Ecol Evol 37, 19–42.

Amir, A., Keselman, D., 1997. Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms. SIAM J. Comput 26 (6), 1656–1669.

Anders, K. L., Nguyet, N. M., Chau, N. V. V., Hung, N. T., Thuy, T. T., Lien, L. B., Farrar, J., Wills, B., Hien, T. T., Simmons, C. P., Jan 2011. Epidemiological factors associated with dengue shock syndrome and mortality in hospitalized dengue patients in ho chi minh city, vietnam. Am J Trop Med Hyg 84 (1), 127–34.

Appel, N., Herian, U., Bartenschlager, R., Jan 2005. Efficient rescue of hepatitis c virus rna replication by trans-complementation with nonstructural protein 5a. J Virol 79 (2), 896–909.

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P., Slowinski, J. B., 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst 33, 707–740.

Balmaseda, A., Standish, K., Mercado, J. C., Matute, J. C., Tellez, Y., Saborío, S., Hammond, S. N., Nuñez, A., Avilés, W., Henn, M. R., Holmes, E. C., Gordon, A., Coloma, J., Kuan, G., Harris, E., Jan 2010. Trends in patterns of dengue transmission over 4 years in a pediatric cohort study in nicaragua. J Infect Dis 201 (1), 5–14.

Baranowski, E., Ruiz-Jarabo, C. M., Domingo, E., May 2001. Evolution of cell recognition by viruses. Science 292 (5519), 1102–5.

Bennett, S. N., Holmes, E. C., Chirivella, M., Rodriguez, D. M., Beltran, M., Vorndam, V., Gubler, D. J., McMillan, W. O., Apr 2006. Molecular evolution of dengue 2 virus in puerto rico: positive selection in the viral envelope accompanies clade reintroduction. J Gen Virol 87 (Pt 4), 885–93.

Blackard, J. T., Sherman, K. E., Feb 2007. Hepatitis c virus coinfection and superinfection. J Infect Dis 195 (4), 519–24.

Bollback, J. P., 2006. Simmap: stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics 7, 88.

Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., Petropoulos, C. J., Nov 2004. Evidence for positive epistasis in hiv-1. Science 306 (5701), 1547–50.

Boom, R., Sol, C. J., Salimans, M. M., Jansen, C. L., Wertheim-van Dillen, P. M., van der Noordaa, J., Mar 1990. Rapid and simple method for purification of nucleic acids. J Clin Microbiol 28 (3), 495–503.

Bowden, S., McCaw, R., White, P. A., Crofts, N., Aitken, C. K., May 2005. Detection of multiple hepatitis c virus genotypes in a cohort of injecting drug users. J Viral Hepat 12 (3), 322–4.

Brown, W. M., George, Jr, M., Wilson, A. C., Apr 1979. Rapid evolution of animal mitochondrial dna. Proc Natl Acad Sci U S A 76 (4), 1967–71.

Brown, W. M., Prager, E. M., Wang, A., Wilson, A. C., 1982. Mitochondrial dna sequences of primates: tempo and mode of evolution. J Mol Evol 18 (4), 225–39.

Bush, R., Bender, C., Subbarao, K., Cox, N., Fitch, W., Dec. 1999. Predicting the evolution of human influenza a. Science 286 (5446), 1921–1925.

Calado, R. A., Rocha, M. R., Parreira, R., Piedade, J., Venenno, T., Esteves, A., Apr 2011. Hepatitis c virus subtypes circulating among intravenous drug users in lisbon, portugal. J Med Virol 83 (4), 608–15.

Cerutti, H., Casas-Mollano, J. A., Aug 2006. On the origin and functions of rna-mediated silencing: from protists to man. Curr Genet 50 (2), 81–99.

Chare, E. R., Gould, E. A., Holmes, E. C., Oct 2003. Phylogenetic analysis reveals a low rate of homologous recombination in negative-sense rna viruses. J Gen Virol 84 (Pt 10), 2691–703.

Chen, R., Holmes, E. C., Dec 2006. Avian influenza virus exhibits rapid evolutionary dynamics. Mol Biol Evol 23 (12), 2336–41.

Chen, R., Holmes, E. C., Jan 2010. Hitchhiking and the population genetic structure of avian influenza virus. J Mol Evol 70 (1), 98–105.

Chinese SARS Molecular Epidemiology Consortium, Mar 2004. Molecular evolution of the sars coronavirus during the course of the sars epidemic in china. Science 303 (5664), 1666–9.

Chua, K. B., Goh, K. J., Wong, K. T., Kamarulzaman, A., Tan, P. S., Ksiazek, T. G., Zaki, S. R., Paul, G., Lam, S. K., Tan, C. T., Oct 1999. Fatal encephalitis due to nipah virus among pig-farmers in malaysia. Lancet 354 (9186), 1257–9.

Claas, E., Osterhaus, A., van Beek, R., De Jong, J., Rimmelzwaan, G., Senne, D., Krauss, S., Shortridge, K., Webster, R., Feb. 1998. Human influenza a h5n1 virus related to a highly pathogenic avian influenza virus. Lancet 351 (9101), 472–477.

Cleaveland, S., Haydon, D. T., Taylor, L., 2007. Overviews of pathogen emergence: which pathogens emerge, when and why? Curr Top Microbiol Immunol 315, 85–111.

Cleaveland, S., Laurenson, M. K., Taylor, L. H., Jul 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos Trans R Soc Lond B Biol Sci 356 (1411), 991–9.

Cobey, S., Koelle, K., Oct 2008. Capturing escape in infectious disease dynamics. Trends Ecol Evol 23 (10), 572–7.

Cole, R., Farach-Colton, M., Hariharan, R., Przytycka, T., Thorup, M., 2000. An o(nlog n) algorithm for the maximum agreement subtree problem for binary trees. SIAM J. Comput 30 (5), 1385–1404.

Colina, R., Casane, D., Vasquez, S., García-Aguirre, L., Chunga, A., Romero, H., Khan, B., Cristina, J., Jan 2004. Evidence of intratypic recombination in natural populations of hepatitis c virus. J Gen Virol 85 (Pt 1), 31–7.

Cologna, R., Armstrong, P. M., Rico-Hesse, R., Jan 2005. Selection for virulent dengue viruses occurs in humans and mosquitoes. J Virol 79 (2), 853–9.

Conzelmann, K. K., 1998. Nonsegmented negative-strand rna viruses: genetics and manipulation of viral genomes. Annu Rev Genet 32, 123–62.

Cranston, K. A., Rannala, B., Aug 2007. Summarizing a posterior distribution of trees using agreement subtrees. Syst Biol 56 (4), 578–90.

Cristina, J., Colina, R., 2006. Evidence of structural genomic region recombination in hepatitis c virus. Virol J 3, 53.

Cui, J., Han, N., Streicker, D., Li, G., Tang, X., Shi, Z., Hu, Z., Zhao, G., Fontanet, A., Guan, Y., Wang, L., Jones, G., Field, H. E., Daszak, P., Zhang, S., Oct 2007. Evolutionary relationships between bat coronaviruses and their hosts. Emerg Infect Dis 13 (10), 1526–32.

Danis, C., Mabrouk, T., Garzon, S., Lemay, G., Mar 1993. Establishment of persistent reovirus infection in sc1 cells: absence of protein synthesis inhibition and increased level of double-stranded rna-activated protein kinase. Virus Res 27 (3), 253–65.

Davies, T. J., Pedersen, A. B., Jul 2008. Phylogeny and geography predict pathogen community similarity in wild primates and humans. Proc Biol Sci 275 (1643), 1695–701.

Diamond, J., Aug 2002. Evolution, consequences and future of plant and animal domestication. Nature 418 (6898), 700–7.

Dobson, A., Carper, E., Feb. 1996. Infectious diseases and human population history - throughout history the establishment of disease has been a side effect of the growth of civilization. Bioscience 46 (2), 115–126.

Dowdle, W., 1999. Influenza a virus recycling revisited. Bulletin of the World Health Organization 77 (10), 820–828.

Drexler, J. F., Gloza-Rausch, F., Glende, J., Corman, V. M., Muth, D., Goettsche, M., Seebens, A., Niedrig, M., Pfefferle, S., Yordanov, S., Zhelyazkov, L., Hermanns, U., Vallo, P., Lukashev, A., Müller, M. A., Deng, H., Herrler, G., Drosten, C., Nov 2010. Genomic characterization of severe acute respiratory syndrome-related coronavirus in european bats and classification of coronaviruses based on partial rna-dependent rna polymerase gene sequences. J Virol 84 (21), 11336–49.

Drummond, A., Pybus, O., Rambaut, A., Forsberg, R., Rodrigo, A., Sep. 2003a. Measurably evolving populations. Trends In Ecology & Evolution 18 (9), 481–488.

Drummond, A., Rambaut, A., Shapiro, B., Pybus, O., May 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22 (5), 1185–1192.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A., May 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol 4 (5), e88.

Drummond, A. J., Pybus, O. G., Rambaut, A., 2003b. Inference of viral evolutionary rates from molecular sequences. Adv Parasitol 54, 331–58.

Drummond, A. J., Rambaut, A., Nov. 2007. Beast: Bayesian evolutionary analysis by sampling trees. Bmc Evolutionary Biology 7, 214.

Drummond, A. J., Suchard, M. A., 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol 8, 114.

Dubois, F., Desenclos, J. C., Mariotte, N., Goudeau, A., Jun 1997. Hepatitis c in a french population-based survey, 1994: seroprevalence, frequency of viremia, genotype distribution, and risk factors. the collaborative study group. Hepatology 25 (6), 1490–6.

Duffy, S., Shackelton, L. A., Holmes, E. C., Apr 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9 (4), 267–76.

Dunham, E. J., Dugan, V. G., Kaser, E. K., Perkins, S. E., Brown, I. H., Holmes, E. C., Taubenberger, J. K., Jun 2009. Different evolutionary trajectories of european avian-like and classical swine h1n1 influenza a viruses. J Virol 83 (11), 5485–94.

Edgar, R. C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32 (5), 1792–7.

Edwards, C. T. T., Holmes, E. C., Pybus, O. G., Wilson, D. J., Viscidi, R. P., Abrams, E. J., Phillips, R. E., Drummond, A. J., Nov 2006. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics 174 (3), 1441–53.

Elena, S. F., Sanjuán, R., 2007. Virus evolution: Insights from an experimental approach. Annu Rev Ecol Evol 38, 27–52.

Endy, T. P., Nisalak, A., Chunsuttitwat, S., Vaughn, D. W., Green, S., Ennis, F. A., Rothman, A. L., Libraty, D. H., Mar 2004. Relationship of preexisting dengue virus (dv) neutralizing antibody levels to viremia and severity of disease in a prospective cohort study of dv infection in thailand. J Infect Dis 189 (6), 990–1000.

Fares, M. A., Holmes, E. C., Jun 2002. A revised evolutionary history of hepatitis b virus (hbv). J Mol Evol 54 (6), 807–14.

Felsenstein, J., Sep 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25 (5), 471–92.

Felsenstein, J., 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J Mol Evol 17 (6), 368–76.

Felsenstein, J., 1985. Phylogenies and the comparative method. Am Nat 125 (1), 1–15.

Ferguson, N., Galvani, A., Bush, R., Mar. 2003. Ecological and immunological determinants of influenza evolution. Nature 422 (6930), 428–433.

Finden, C. R., Gordon, A. D., 1985. Obtaining common pruned trees. J Classif 2 (1), 255–276.

Fitch, W. M., Bush, R. M., Bender, C. A., Cox, N. J., Jul 1997. Long term trends in the evolution of h(3) ha1 human influenza type a. Proc Natl Acad Sci U S A 94 (15), 7712–8.

Fitch, W. M., Leiter, J. M., Li, X. Q., Palese, P., May 1991. Positive darwinian evolution in human influenza a viruses. Proc Natl Acad Sci U S A 88 (10), 4270–4.

Foster, J. E., Bennett, S. N., Vaughan, H., Vorndam, V., McMillan, W. O., Carrington, C. V. F., Feb 2003. Molecular evolution and phylogeny of dengue type 4 virus in the caribbean. Virology 306 (1), 126–34.

Francesconi, P., Yoti, Z., Declich, S., Onek, P. A., Fabiani, M., Olango, J., Andraghetti, R., Rollin, P. E., Opira, C., Greco, D., Salmaso, S., Nov 2003. Ebola hemorrhagic fever transmission and risk factors of contacts, uganda. Emerg Infect Dis 9 (11), 1430–7.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., WHO Rapid Pandemic Assessment Collaboration, Jun 2009. Pandemic potential of a strain of influenza a (h1n1): early findings. Science 324 (5934), 1557–61.

Gouilh, M. A., Puechmaille, S. J., Gonzalez, J.-P., Teeling, E., Kittayapong, P., Manuguerra, J.-C., Jul 2011. Sars-coronavirus ancestor's foot-prints in south-east asian bat colonies and the refuge theory. Infect Genet Evol.

Graham, R. L., Baric, R. S., Apr 2010. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. J Virol 84 (7), 3134–46.

Gray, R. R., Parker, J., Lemey, P., Salemi, M., Katzourakis, A., Pybus, O. G., May 2011. The mode and tempo of hepatitis c virus evolution within and among hosts. Bmc Evolutionary Biology 11, 131.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., Holmes, E. C., Jan 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303 (5656), 327–32.

Guan, Y., Peiris, J., Lipatov, A., Ellis, T., Dyrting, K., Krauss, S., Zhang, L., Webster, R., Shortridge, K., Jun. 2002. Emergence of multiple genotypes of h5n1 avian influenza viruses in hong kong sar. Proc Natl Acad Sci U S A 99 (13), 8950–8955.

Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S. M., Poon, L. L. M., Oct 2003. Isolation and characterization of viruses related to the sars coronavirus from animals in southern china. Science 302 (5643), 276–8.

Gubler, D. J., Aug 1987. Dengue and dengue hemorrhagic fever in the americas. P R Health Sci J 6 (2), 107–11.

Gubler, D. J., Jul 1998. Dengue and dengue hemorrhagic fever. Clin Microbiol Rev 11 (3), 480–96.

Gubler, D. J., Meltzer, M., 1999. Impact of dengue/dengue hemorrhagic fever on the developing world. Adv Virus Res 53, 35–70.

Gubler, D. J., Trent, D. W., Dec 1993. Emergence of epidemic dengue/dengue hemorrhagic fever as a public health problem in the americas. Infect Agents Dis 2 (6), 383–93.

Guzmán, M. G., Kourí, G., Jan 2002. Dengue: an update. Lancet Infect Dis 2 (1), 33–42.

Guzmán, M. G., Kouri, G., Valdes, L., Bravo, J., Alvarez, M., Vazques, S., Delgado, I., Halstead, S. B., Nov 2000. Epidemiologic studies on dengue in santiago de cuba, 1997. Am J Epidemiol 152 (9), 793–9; discussion 804.

Hang, V. T. T., Holmes, E. C., Duong, V., Nguyen, T. Q., Tran, T. H., Quail, M., Churcher, C., Parkhill, J., Cardosa, J., Farrar, J., Wills, B., Lennon, N. J., Birren, B. W., Buchy, P., Henn, M. R., Simmons, C. P., 2010. Emergence of the asian 1 genotype of dengue virus serotype 2 in viet nam: in vivo fitness advantage and lineage replacement in south-east asia. PLoS Negl Trop Dis 4 (7), e757.

Harrington, L. C., Scott, T. W., Lerdthusnee, K., Coleman, R. C., Costero, A., Clark, G. G., Jones, J. J., Kitthawee, S., Kittayapong, P., Sithiprasasna, R., Edman, J. D., Feb 2005. Dispersal of the dengue vector aedes aegypti within and between rural communities. Am J Trop Med Hyg 72 (2), 209–20.

Harvey, P. H., Pagel, M. D., 1991. The comparative method in evolutionary biology. OXFORD UNIV PRESS.

Herlihy, K. J., Graham, J. P., Kumpf, R., Patick, A. K., Duggal, R., Shi, S. T., Oct 2008. Development of intergenotypic chimeric replicons to determine the broad-spectrum antiviral activities of hepatitis c virus polymerase inhibitors. Antimicrob Agents Chemother 52 (10), 3523–31.

Herring, B. L., Page-Shafer, K., Tobler, L. H., Delwart, E. L., Oct 2004. Frequent hepatitis c virus superinfection in injection drug users. J Infect Dis 190 (8), 1396–403.

Holder, M., Lewis, P. O., Apr 2003. Phylogeny estimation: traditional and bayesian approaches. Nat Rev Genet 4 (4), 275–84.

Holmes, E. C., Dec 2003a. Error thresholds and the constraints to rna virus evolution. Trends Microbiol 11 (12), 543–6.

Holmes, E. C., Apr 2003b. Molecular clocks and the puzzle of rna virus origins. J Virol 77 (7), 3893–7.

Holmes, E. C., Mar 2006. The evolution of viral emergence. Proc Natl Acad Sci U S A 103 (13), 4803–4.

Holmes, E. C., 2008. Evolutionary history and phylogeography of human viruses. Annu Rev Microbiol 62, 307–28.

Holmes, E. C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B. T., Salzberg, S. L., Fraser, C. M., Lipman, D. J., Taubenberger, J. K., Sep 2005. Whole-genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment among recent h3n2 viruses. PLoS Biol 3 (9), e300.

Holmes, E. C., Rambaut, A., Jul 2004. Viral evolution and the emergence of sars coronavirus. Philos Trans R Soc Lond B Biol Sci 359 (1447), 1059–65.

Holmes, E. C., Twiddy, S. S., May 2003. The origin, emergence and evolutionary genetics of dengue virus. Infect Genet Evol 3 (1), 19–28.

Hon, C.-C., Lam, T.-Y., Shi, Z.-L., Drummond, A. J., Yip, C.-W., Zeng, F., Lam, P.-Y., Leung, F. C.-C., Feb 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (sars)-like coronavirus and its implications on the direct ancestor of sars coronavirus. J Virol 82 (4), 1819–26.

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., Ikuta, K., Jern, P., Gojobori, T., Coffin, J. M., Tomonaga, K., Jan 2010. Endogenous non-retroviral rna virus elements in mammalian genomes. Nature 463 (7277), 84–7.

Hu, W. S., Temin, H. M., Feb 1990. Genetic consequences of packaging two rna genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc Natl Acad Sci U S A 87 (4), 1556–60.

Huelsenbeck, J. P., Nielsen, R., Bollback, J. P., Apr 2003. Stochastic mapping of morphological characters. Syst Biol 52 (2), 131–58.

Huestis, D., Apr. 2002. Russia's national research center for hematology: its role in the development of blood banking. Transfusion 42 (4), 490–494.

Initative, P. D. V., November 2009. Global burden of dengue.
URL `http://www.pdvi.org/about_dengue/GBD.asp`

Iyer, L. M., Balaji, S., Koonin, E. V., Aravind, L., Apr 2006. Evolutionary genomics of nucleo-cytoplasmic large dna viruses. Virus Res 117 (1), 156–84.

Jarman, R. G., Holmes, E. C., Rodpradit, P., Klungthong, C., Gibbons, R. V., Nisalak, A., Rothman, A. L., Libraty, D. H., Ennis, F. A., Mammen, Jr, M. P., Endy, T. P., Jun 2008. Microevolution of dengue viruses circulating among primary school children in kamphaeng phet, thailand. J Virol 82 (11), 5494–500.

Jenkins, G. M., Rambaut, A., Pybus, O. G., Holmes, E. C., Feb 2002. Rates of molecular evolution in rna viruses: a quantitative phylogenetic analysis. J Mol Evol 54 (2), 156–65.

Kageyama, S., Agdamag, D. M., Alesna, E. T., Leano, P. S., Heredia, A. M. L., Abellanosa-Tac-An, I. P., Jereza, L. D., Tanimoto, T., Yamamura, J.-i., Ichimura, H., Nov. 2006. A natural inter-genotypic (2b/1b) recombinant of hepatitis c virus in the philippines. J Med Virol 78 (11), 1423–1428.

Kalinina, O., Norder, H., Mukomolov, S., Magnius, L., Apr. 2002. A natural intergenotypic recombinant of hepatitis c virus identified in st. petersburg. J Virol 76 (8), 4034–4043.

Kalinina, O., Norder, H., Vetrov, T., Zhdanov, K., Barzunova, M., Plotnikova, V., Mukomolov, S., Magnius, L., Nov. 2001. Shift in predominating subtype of hcv from 1b to 3a in st. petersburg mediated by increase in injecting drug use. J Med Virol 65 (3), 517–524.

Kan, B., Wang, M., Jing, H., Xu, H., Jiang, X., Yan, M., Liang, W., Zheng, H., Wan, K., Liu, Q., Cui, B., Xu, Y., Zhang, E., Wang, H., Ye, J., Li, G., Li, M., Cui, Z., Qi, X., Chen, K., Du, L., Gao, K., Zhao, Y.-T., Zou, X.-Z., Feng, Y.-J., Gao, Y.-F., Hai, R., Yu, D., Guan, Y., Xu, J., Sep 2005. Molecular evolution analysis and geographic investigation of severe acute respiratory syndrome coronavirus-like virus in palm civets at an animal market and on farms. J Virol 79 (18), 11892–900.

Kapoor, A., Simmonds, P., Gerold, G., Qaisar, N., Jain, K., Henriquez, J. A., Firth, C., Hirschberg, D. L., Rice, C. M., Shields, S., Lipkin, W. I., May 2011. Characterization of a canine homolog of hepatitis c virus. Proc Natl Acad Sci U S A.

Katzourakis, A., Gifford, R. J., Nov 2010. Endogenous viral elements in animal genomes. PLoS Genet 6 (11), e1001191.

Kida, H., Ito, T., Yasuda, J., Shimizu, Y., Itakura, C., Shortridge, K. F., Kawaoka, Y., Webster, R. G., Sep 1994. Potential for transmission of avian influenza viruses to pigs. J Gen Virol 75 (Pt 9), 2183–8.

Kingman, J. F. C., 1982. The coalescent. Stochastic Processes and their Applications 13 (3), 235 – 248.

Koelle, K., Cobey, S., Grenfell, B., Pascual, M., Dec 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. Science 314 (5807), 1898–903.

Koonin, E. V., Senkevich, T. G., Dolja, V. V., 2006. The ancient virus world and evolution of cells. Biol Direct 1, 29.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., Bhattacharya, T., Jun 2000. Timing the ancestor of the hiv-1 pandemic strains. Science 288 (5472), 1789–96.

Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H., Frost, S. D. W., Dec 2006. Gard: a genetic algorithm for recombination detection. Bioinformatics 22 (24), 3096–8.

Kuiken, C., Simmonds, P., 2009. Nomenclature and numbering of the hepatitis c virus. Methods Mol Biol 510, 33–53.

Kuiken, C., Yusim, K., Boykin, L., Richardson, R., Feb 2005. The los alamos hepatitis c sequence database. Bioinformatics 21 (3), 379–84.

Kuiken, T., Holmes, E. C., McCauley, J., Rimmelzwaan, G. F., Williams, C. S., Grenfell, B. T., Apr 2006. Host species barriers to influenza virus infections. Science 312 (5772), 394–7.

Kurbanov, F., Tanaka, Y., Avazova, D., Khan, A., Sugauchi, F., Kan, N., Kurbanova-Khudayberganova, D., Khikmatullaeva, A., Musabaev, E., Mizokami, M., Apr 2008a. Detection of hepatitis c virus natural recombinant rf1_2k/1b strain among intravenous drug users in uzbekistan. Hepatol Res 38 (5), 457–464.

Kurbanov, F., Tanaka, Y., Chub, E., Maruyama, I., Azlarova, A., Kamitsukasa, H., Ohno, T., Bonetto, S., Moreau, I., Fanning, L. J., Legrand-Abravanel, F., Izopet, J., Naoumov, N., Shimada, T., Netesov, S., Mizokami, M., Nov 2008b. Molecular epidemiology and interferon susceptibility of the natural recombinant hepatitis c virus strain rf1_2k/1b. J Infect Dis 198 (10), 1448–56.

Kurbanov, F., Tanaka, Y., Fujiwara, K., Sugauchi, F., Mbanya, D., Zekeng, L., Ndembi, N., Ngansop, C., Kaptue, L., Miura, T., Ido, E., Hayami, M., Ichimura, H., Mizokami, M., Jul. 2005. A new subtype (subgenotype) ac (a3) of hepatitis b virus and recombination between genotypes a and e in cameroon. J Gen Virol 86, 2047–2056.

Lai, M. M. C., Perlman, S., Anderson, L. J., 2007. Fields Virology: Coronaviridae, 5th Edition. Vol. 1. Philadelphia: Lippincott Williams and Wilkins.

Lam, T.-Y., Hon, C.-C., Wang, Z., Hui, R. K.-H., Zeng, F., Leung, F. C.-C., Feb. 2008. Evolutionary analyses of european h1n2 swine influenza a virus by placing timestamps on the multiple reassortment events. Virus Research 131 (2), 271–278.

Lanciotti, R. S., Gubler, D. J., Trent, D. W., Sep 1997. Molecular evolution and phylogeny of dengue-4 viruses. J Gen Virol 78 (Pt 9), 2279–84.

Larget, B., Simon, D. L., 1999. Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. Mol Biol Evol 16, 750–759.

Lau, S. K. P., Li, K. S. M., Huang, Y., Shek, C.-T., Tse, H., Wang, M., Choi, G. K. Y., Xu, H., Lam, C. S. F., Guo, R., Chan, K.-H., Zheng, B.-J., Woo, P. C. Y., Yuen, K.-Y., Mar 2010. Ecoepidemiology and complete genome comparison of different strains of severe acute respiratory syndrome-related rhinolophus bat coronavirus in china reveal bats as a reservoir for acute, self-limiting infection that allows recombination events. J Virol 84 (6), 2808–19.

Lau, S. K. P., Woo, P. C. Y., Li, K. S. M., Huang, Y., Tsoi, H.-W., Wong, B. H. L., Wong, S. S. Y., Leung, S.-Y., Chan, K.-H., Yuen, K.-Y., Sep 2005. Severe acute respiratory syndrome coronavirus-like virus in chinese horseshoe bats. Proc Natl Acad Sci U S A 102 (39), 14040–5.

Lee, C.-M., Hung, L.-J., Chang, M.-S., Shen, C.-B., Tang, C.-Y., June 2005. An improved algorithm for the maximum agreement subtree problem. Inf. Process. Lett. 94, 211–216.

Lee, Y.-M., Lin, H.-J., Chen, Y.-J., Lee, C.-M., Wang, S.-F., Chang, K.-Y., Chen, T.-L., Liu, H.-F., Chen, Y.-M. A., Jan. 2010. Molecular epidemiology of hcv genotypes among injection drug users in taiwan: Full-length sequences of two new subtype 6w strains and a recombinant form_2b6w. J Med Virol 82 (1), 57–68.

Legrand-Abravanel, F., Claudinon, J., Nicot, F., Dubois, M., Chapuy-Regaud, S., Sandres-Saune, K., Pasquier, C., Izopet, J., Apr. 2007. New natural intergenotypic (2/5) recombinant of hepatitis c virus. J Virol 81 (8), 4357–4362.

Leitner, T., Korber, B., Daniels, M., Calef, C., Foley, B., 2005. Hiv-1 subtype and circulating recombinant form (crf) reference sequences.
URL `http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/LEITNER2005/leitner.html`

Lemey, P., Rambaut, A., Drummond, A. J., Suchard, M. A., Sep 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol 5 (9), e1000520.

Lemey, P., Rambaut, A., Pybus, O. G., 2006. Hiv evolutionary dynamics within and among hosts. AIDS Rev 8 (3), 125–40.

Lemey, P., Rambaut, A., Welch, J. J., Suchard, M. A., Aug 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol Biol Evol 27 (8), 1877–85.

Lemon, S., Brown, E., 1995. Hepatitis C Virus, 4th Edition. No. 1474-1486 in Principle and Practice of Infectious Disease. Churchill Livingstone, New York.

Leroy, E. M., Kumulungui, B., Pourrut, X., Rouquet, P., Hassanin, A., Yaba, P., Délicat, A., Paweska, J. T., Gonzalez, J.-P., Swanepoel, R., Dec 2005. Fruit bats as reservoirs of ebola virus. Nature 438 (7068), 575–6.

Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., Somasundaran, M., Sullivan, J. L., Luzuriaga, K., Greenough, T. C., Choe, H., Farzan, M., Nov 2003. Angiotensin-converting enzyme 2 is a functional receptor for the sars coronavirus. Nature 426 (6965), 450–4.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J. H., Wang, H., Crameri, G., Hu, Z., Zhang, H., Zhang, J., McEachern, J., Field, H., Daszak, P., Eaton, B. T., Zhang, S., Wang, L.-F., Oct 2005. Bats are natural reservoirs of sars-like coronaviruses. Science 310 (5748), 676–9.

Li, W., Wong, S.-K., Li, F., Kuhn, J. H., Huang, I.-C., Choe, H., Farzan, M., May 2006. Animal origins of the severe acute respiratory syndrome coronavirus: insight from ace2-s-protein interactions. J Virol 80 (9), 4211–9.

Libraty, D. H., Endy, T. P., Houng, H.-S. H., Green, S., Kalayanarooj, S., Suntayakorn, S., Chansiriwongs, W., Vaughn, D. W., Nisalak, A., Ennis, F. A., Rothman, A. L., May 2002. Differing influences of virus burden and immune activation on disease severity in secondary dengue-3 virus infections. J Infect Dis 185 (9), 1213–21.

Mackenzie, J. S., 1999. Emerging viral diseases: an australian perspective. Emerg Infect Dis 5 (1), 1–8.

Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S. Y. W., Shapiro, B., Pybus, O. G., Allain, J.-P., Hatzakis, A., Dec. 2009. The global spread of hepatitis c virus 1a and 1b: A phylodynamic and phylogeographic analysis. Plos Medicine 6 (12), e1000198.

Mammen, M. P., Pimgate, C., Koenraadt, C. J. M., Rothman, A. L., Aldstadt, J., Nisalak, A., Jarman, R. G., Jones, J. W., Srikiatkhachorn, A., Ypil-Butac, C. A., Getis, A., Thammapalo, S., Morrison, A. C., Libraty, D. H., Green, S., Scott, T. W., Nov 2008. Spatial and temporal clustering of dengue virus transmission in thai villages. PLoS Med 5 (11), e205.

Markov, P. V., Pepin, J., Frost, E., Deslandes, S., Labbe, A.-C., Pybus, O. G., Sep. 2009. Phylogeography and molecular epidemiology of hepatitis c virus genotype 2 in africa. J Gen Virol 90, 2086–2096.

Marra, M. A., Jones, S. J. M., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S. N., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., Pandoh, P. K., Petrescu, A. S., Robertson, A. G., Schein, J. E., Siddiqui, A., Smailus, D. E., Stott, J. M., Yang, G. S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T. F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G. A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R. C., Krajden, M., Petric, M., Skowronski, D. M., Upton, C., Roper, R. L., May 2003. The genome sequence of the sars-associated coronavirus. Science 300 (5624), 1399–404.

Martina, B. E. E., Haagmans, B. L., Kuiken, T., Fouchier, R. A. M., Rimmelzwaan, G. F., Van Amerongen, G., Peiris, J. S. M., Lim, W., Osterhaus, A. D. M. E., Oct 2003. Virology: Sars virus infection of cats and ferrets. Nature 425 (6961), 915.

Matusevich, M., 2009. Black in the u.s.s.r.: Africans, african american, and the soviet society. Transition (100), 56–75.

Mau, B., Newton, M. A., 1997. Phylogenetic inference for binary data on dendrograms using markov chain monte carol. J Comput Graph Stat 6, 122–131.

Mau, B., Newton, M. A., Larget, B., Mar 1999. Bayesian phylogenetic inference via markov chain monte carlo methods. Biometrics 55 (1), 1–12.

McElroy, K. L., Santiago, G. A., Lennon, N. J., Birren, B. W., Henn, M. R., Muñoz-Jordán, J. L., Jan 2011. Endurance, refuge, and reemergence of dengue virus type 2, puerto rico, 1986-2007. Emerg Infect Dis 17 (1), 64–71.

Messer, W. B., Gubler, D. J., Harris, E., Sivananthan, K., de Silva, A. M., Jul 2003. Emergence and global spread of a dengue serotype 3, subtype iii virus. Emerg Infect Dis 9 (7), 800–9.

Minin, V. N., Bloomquist, E. W., Suchard, M. A., Jul. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol 25 (7), 1459–1471.

Minin, V. N., Suchard, M. A., Mar 2008. Counting labeled transitions in continuous-time markov models of evolution. J Math Biol 56 (3), 391–412.

Morel, V., Descamps, V., Francois, C., Fournier, C., Brochot, E., Capron, D., Duverlie, G., Castelain, S., Apr. 2010. Emergence of a genomic variant of the recombinant 2k/1b strain during a mixed hepatitis c infection: A case report. J Clin Virol 47 (4), 382–386.

Moreno, P., Alvarez, M., Lopez, L., Moratorio, G., Casane, D., Castells, M., Castro, S., Cristina, J., Colina, R., Nov. 2009. Evidence of recombination in hepatitis c virus populations infecting a hemophiliac patient. Virology Journal 6, 203.

Morrison, A. C., Minnick, S. L., Rocha, C., Forshey, B. M., Stoddard, S. T., Getis, A., Focks, D. A., Russell, K. L., Olson, J. G., Blair, P. J., Watts, D. M., Sihuincha, M., Scott, T. W., Kochel, T. J., 2010. Epidemiology of dengue virus in iquitos, peru 1999 to 2005: interepidemic and epidemic patterns of transmission. PLoS Negl Trop Dis 4 (5), e670.

Morse, S. S., 1995. Factors in the emergence of infectious diseases. Emerg Infect Dis 1 (1), 7–15.

Murphy, D. G., Willems, B., Deschenes, M., Hilzenrat, N., Mousseau, R., Sabbah, S., Apr. 2007. Use of sequence analysis of the ns5b region for routine genotyping of hepatitis c virus with reference to c/e1 and 5 ' untranslated region sequences. J Clin Microbiol 45 (4), 1102–1112.

Myat Thu, H., Lowry, K., Jiang, L., Hlaing, T., Holmes, E. C., Aaskov, J., Jun 2005. Lineage extinction and replacement in dengue type 1 virus populations are due to stochastic events rather than to natural selection. Virology 336 (2), 163–72.

Nagao, Y., Koelle, K., Feb 2008. Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever. Proc Natl Acad Sci U S A 105 (6), 2238–43.

Nagarajan, N., Kingsford, C., Mar 2011. Giraf: robust, computational identification of influenza reassortments via graph mining. Nucleic Acids Res 39 (6), e34.

Nakajima, K., Desselberger, U., Palese, P., Jul 1978. Recent human influenza a (h1n1) viruses are closely related genetically to strains isolated in 1950. Nature 274 (5669), 334–9.

Nelson, M. I., Edelman, L., Spiro, D. J., Boyne, A. R., Bera, J., Halpin, R., Sengamalay, N., Ghedin, E., Miller, M. A., Simonsen, L., Viboud, C., Holmes, E. C., 2008a. Molecular epidemiology of a/h3n2 and a/h1n1 influenza virus during a single epidemic season in the united states. PLoS Pathog 4 (8), e1000133.

Nelson, M. I., Lemey, P., Tan, Y., Vincent, A., Lam, T. T.-Y., Detmer, S., Viboud, C., Suchard, M. A., Rambaut, A., Holmes, E. C., Gramer, M., Jun 2011. Spatial dynamics of human-origin h1 influenza a virus in north american swine. PLoS Pathog 7 (6), e1002077.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., Holmes, E. C., Sep 2007. Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. PLoS Pathog 3 (9), 1220–8.

Nelson, M. I., Simonsen, L., Viboud, C., Miller, M. A., Taylor, J., George, K. S., Griesemer, S. B., Ghedin, E., Ghedi, E., Sengamalay, N. A., Spiro, D. J., Volkov, I., Grenfell, B. T., Lipman, D. J., Taubenberger, J. K., Holmes, E. C., Dec 2006. Stochastic processes are key determinants of short-term evolution in influenza a virus. PLoS Pathog 2 (12), e125.

Nelson, M. I., Viboud, C., Simonsen, L., Bennett, R. T., Griesemer, S. B., St George, K., Taylor, J., Spiro, D. J., Sengamalay, N. A., Ghedin, E., Taubenberger, J. K., Holmes, E. C., Feb 2008b. Multiple reassortment events in the evolutionary history of h1n1 influenza a virus since 1918. PLoS Pathog 4 (2), e1000012.

Newton, M. A., Raftery, A. E., 1994. Approximate bayesian inference with the weighted likelihood bootstrap. J Royal Stat Society-Series B 56 (1), 3–48.

Nielsen, R., Sep 2001. Mutations as missing data: inferences on the ages and distributions of nonsynonymous and synonymous mutations. Genetics 159 (1), 401–11.

Noppornpanth, S., Lien, T. X., Poovorawan, Y., Smits, S. L., Osterhaus, A. D. M. E., Haagmans, B. L., Aug. 2006. Identification of a naturally occurring recombinant genotype 2/6 hepatitis c virus. J Virol 80 (15), 7569–7577.

O'Brien, J. D., Minin, V. N., Suchard, M. A., Apr 2009. Learning to count: robust estimates for labeled distances between molecular sequences. Mol Biol Evol 26 (4), 801–14.

O'Brien, J. D., She, Z.-S., Suchard, M. A., 2008. Dating the time of viral subtype divergence. BMC Evol Biol 8, 172.

Osburn, W. O., Fisher, B. E., Dowd, K. A., Urban, G., Liu, L., Ray, S. C., Thomas, D. L., Cox, A. L., Jan 2010. Spontaneous control of primary hepatitis c virus infection and immunity against persistent reinfection. Gastroenterology 138 (1), 315–24.

Pagel, M., Oct 1999. Inferring the historical patterns of biological evolution. Nature 401 (6756), 877–84.

Palese, P., Dec 2004. Influenza: old and new threats. Nat Med 10 (12 Suppl), S82–7.

Pawlotsky, J. M., Tsakiris, L., Roudot-Thoraval, F., Pellet, C., Stuyver, L., Duval, J., Dhumeaux, D., Jun 1995. Relationship between hepatitis c virus genotypes and sources of infection in patients with chronic hepatitis c. J Infect Dis 171 (6), 1607–10.

Pearce-Duvet, J. M. C., Aug 2006. The origin of human pathogens: evaluating the role of agriculture and domestic animals in the evolution of human disease. Biol Rev Camb Philos Soc 81 (3), 369–82.

Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Guan, Y., Yam, L. Y. C., Lim, W., Nicholls, J., Yee, W. K. S., Yan, W. W., Cheung, M. T., Cheng, V. C. C., Chan, K. H., Tsang, D. N. C., Yung, R. W. H., Ng, T. K., Yuen, K. Y., SARS study group, Apr 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361 (9366), 1319–25.

Pol, S., Thiers, V., Nousbaum, J. B., Legendre, C., Berthelot, P., Kreis, H., Brechot, C., Feb 1995. The changing relative prevalence of hepatitis c virus genotypes: evidence in hemodialyzed patients and kidney recipients. Gastroenterology 108 (2), 581–3.

Poon, L. L. M., Chu, D. K. W., Chan, K. H., Wong, O. K., Ellis, T. M., Leung, Y. H. C., Lau, S. K. P., Woo, P. C. Y., Suen, K. Y., Yuen, K. Y., Guan, Y., Peiris, J. S. M., Feb 2005. Identification of a novel coronavirus in bats. J Virol 79 (4), 2001–9.

Posada, D., Crandall, K. A., Mar 2002. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 54 (3), 396–402.

Posada, D., Crandall, K. A., Holmes, E. C., 2002. Recombination in evolutionary genomics. Annu Rev Genet 36, 75–97.

Prangishvili, D., Garrett, R. A., Koonin, E. V., Apr 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res 117 (1), 52–67.

Pybus, O., Charleston, M., Gupta, S., Rambaut, A., Holmes, E., Harvey, P., Jun. 2001. The epidemic behavior of the hepatitis c virus. Science 292 (5525), 2323–2325.

Pybus, O., Drummond, A., Nakano, T., Robertson, B., Rambaut, A., Mar. 2003. The epidemiology and iatrogenic transmission of hepatitis c virus in egypt: A bayesian coalescent approach. Mol Biol Evol 20 (3), 381–387.

Pybus, O. G., Barnes, E., Taggart, R., Lemey, P., Markov, P. V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I. S., Lu, L., Newton, P. N., Klenerman, P., Jan. 2009. Genetic history of hepatitis c virus in east asia. J Virol 83 (2), 1071–1082.

Pybus, O. G., Cochrane, A., Holmes, E. C., Simmonds, P., Mar 2005. The hepatitis c virus epidemic among injecting drug users. Infect Genet Evol 5 (2), 131–9.

Pybus, O. G., Markov, P. V., Wu, A., Tatem, A. J., Jul. 2007a. Investigating the endemic transmission of the hepatitis c virus. International Journal For Parasitology 37 (8-9), 839–849.

Pybus, O. G., Rambaut, A., Aug. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews Genetics 10 (8), 540–550.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J., Holmes, E. C., Mar 2007b. Phylogenetic evidence for deleterious mutation load in rna viruses and its contribution to viral evolution. Mol Biol Evol 24 (3), 845–52.

Pybus, O. G., Rambaut, A., Harvey, P. H., 2000. An integrated framework for the inference of viral population history from reconstructed genelogies. Genetics 155, 1429–1437.

Qian, K. P., Natov, S. N., Pereira, B. J., Lau, J. Y., Mar 2000. Hepatitis c virus mixed genotype infection in patients on haemodialysis. J Viral Hepat 7 (2), 153–60.

Rabaa, M. A., Ty Hang, V. T., Wills, B., Farrar, J., Simmons, C. P., Holmes, E. C., 2010. Phylogeography of recently emerged denv-2 in southern viet nam. PLoS Negl Trop Dis 4 (7), e766.

Raghwani, J., Rambaut, A., Holmes, E. C., Hang, V. T., Hien, T. T., Farrar, J., Wills, B., Lennon, N. J., Birren, B. W., Henn, M. R., Simmons, C. P., Jun 2011. Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. PLoS Pathog 7 (6), e1002064.

Raghwani, J., Thomas, X. V., Koekkoek, S. M., Schinkel, J., Molenkamp, R., van de Laar, T. J., Takebe, Y., Tanaka, Y., Mizokami, M., Rambaut, A., Pybus, O. G., Nov 2012. The origin and evolution of the unique hcv circulating recombinant form 2k/1b. J Virol 86 (4), 2212–2220.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., Holmes, E. C., May 2008. The genomic and epidemiological dynamics of human influenza a virus. Nature 453 (7195), 615–U2.

Reiter, P., Amador, M. A., Anderson, R. A., Clark, G. G., Feb 1995. Short report: dispersal of aedes aegypti in an urban area after blood feeding as demonstrated by rubidium-marked eggs. Am J Trop Med Hyg 52 (2), 177–9.

Ren, W., Li, W., Yu, M., Hao, P., Zhang, Y., Zhou, P., Zhang, S., Zhao, G., Zhong, Y., Wang, S., Wang, L.-F., Shi, Z., Nov 2006. Full-length genome sequences of two sars-like coronaviruses in horseshoe bats and genetic variation analysis. J Gen Virol 87 (Pt 11), 3355–9.

Ren, W., Qu, X., Li, W., Han, Z., Yu, M., Zhou, P., Zhang, S.-Y., Wang, L.-F., Deng, H., Shi, Z., Feb 2008. Difference in receptor usage between severe acute respiratory syndrome (sars) coronavirus and sars-like coronavirus of bat origin. J Virol 82 (4), 1899–907.

Rico-Hesse, R., Harrison, L. M., Salas, R. A., Tovar, D., Nisalak, A., Ramos, C., Boshell, J., de Mesa, M. T., Nogueira, R. M., da Rosa, A. T., Apr 1997. Origins of dengue type 2 viruses associated with increased pathogenicity in the americas. Virology 230 (2), 244–51.

Ristic, N., Zukurov, J., Alkmim, W., Diaz, R. S., Janini, L. M., Chin, M. P. S., Mar. 2011. Analysis of the origin and evolutionary history of hiv-1 crf28_bf and crf29_bf reveals a decreasing prevalence in the aids epidemic of brazil. PLoS One 6 (3), e17485.

Robertson, D., Sharp, P., McCutchan, F., Hahn, B., Mar. 1995. Recombination in hiv-1. Nature 374 (6518), 124–126.

Robinson, D. F., Foulds, L. R., 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53 (1-2), 131–147.

Rodrigo, A. G., Felsenstein, J., 1999. Coalescent approaches to HIV population genetics, in The Evolution of HIV. John Hopkins University Press.

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Peñaranda, S., Bankamp, B., Maher, K., Chen, M.-H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J. L., Chen, Q., Wang, D., Erdman, D. D., Peret, T. C. T., Burns, C., Ksiazek, T. G., Rollin, P. E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Günther, S., Osterhaus, A. D. M. E., Drosten, C., Pallansch, M. A., Anderson, L. J., Bellini, W. J., May 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300 (5624), 1394–9.

Rudicell, R. S., Holland Jones, J., Wroblewski, E. E., Learn, G. H., Li, Y., Robertson, J. D., Greengrass, E., Grossmann, F., Kamenya, S., Pintea, L., Mjungu, D. C., Lonsdorf, E. V., Mosser, A., Lehman, C., Collins, D. A., Keele, B. F., Goodall, J., Hahn, B. H., Pusey, A. E., Wilson, M. L., 2010. Impact of simian immunodeficiency virus infection on chimpanzee population dynamics. PLoS Pathog 6 (9).

Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., Smith, D. J., Apr. 2008. The global circulation of seasonal influenza a (h3n2) viruses. Science 320 (5874), 340–346.

Russell, R., Webb, C., Williams, C., Ritchie, S., Dec. 2005. Mark-release-recapture study to measure dispersal of the mosquito aedes aegypti in cairns, queensland, australia. Medical and Veterinary Entomology 19 (4), 451–457.

Sanjuán, R., Cuevas, J. M., Moya, A., Elena, S. F., Jul 2005. Epistasis and the adaptability of an rna virus. Genetics 170 (3), 1001–8.

Scholtissek, C., Dec 1987. Molecular aspects of the epidemiology of virus disease. Experientia 43 (11-12), 1197–201.

Scholtissek, C., von Hoyningen, V., Rott, R., Sep 1978. Genetic relatedness between the new 1977 epidemic strains (h1n1) of influenza and human influenza strains isolated between 1947 and 1957 (h1n1). Virology 89 (2), 613–7.

Schrag, S. J., Wiener, P., Aug 1995. Emerging infectious disease: what are the relative roles of ecology and evolution? Trends Ecol Evol 10 (8), 319–24.

Schreiber, G. B., Busch, M. P., Kleinman, S. H., Korelitz, J. J., Jun 1996. The risk of transfusion-transmitted viral infections. the retrovirus epidemiology donor study. N Engl J Med 334 (26), 1685–90.

Schreiber, M. J., Holmes, E. C., Ong, S. H., Soh, H. S. H., Liu, W., Tanner, L., Aw, P. P. K., Tan, H. C., Ng, L. C., Leo, Y. S., Low, J. G. H., Ong, A., Ooi, E. E., Vasudevan, S. G., Hibberd, M. L., May 2009. Genomic epidemiology of a dengue virus epidemic in urban singapore. J Virol 83 (9), 4163–73.

Schröter, M., Feucht, H.-H., Zöllner, B., Schäfer, P., Laufs, R., Jul 2003. Multiple infections with different hcv genotypes: prevalence and clinical impact. J Clin Virol 27 (2), 200–4.

Scott, T. W., Amerasinghe, P. H., Morrison, A. C., Lorenz, L. H., Clark, G. G., Strickman, D., Kittayapong, P., Edman, J. D., Jan 2000. Longitudinal studies of aedes aegypti (diptera: Culicidae) in thailand and puerto rico: blood feeding frequency. J Med Entomol 37 (1), 89–101.

Scott, T. W., Morrison, A. C., 2003. Aedes aegypti density and the risk of dengue-virus transmission. In: Takken, W., Scott, T. (Eds.), ECOLOGICAL ASPECTS FOR APPLICATION OF GENETICALLY MODIFIED MOSQUITOES. Vol. 2 of WAGENINGEN UR FRONTIS SERIES. Frontis, SPRINGER, PO BOX 17, 3300 AA DORDRECHT, NETHERLANDS, pp. 187–206, workshop on Ecological Aspects for Application of Genetically Modified Mosquitoes, WAGENINGEN, NETHERLANDS, JUN, 2002.

Seeff, L. B., Miller, R. N., Rabkin, C. S., Buskell-Bales, Z., Straley-Eason, K. D., Smoak, B. L., Johnson, L. D., Lee, S. R., Kaplan, E. L., Jan 2000. 45-year follow-up of hepatitis c virus infection in healthy young adults. Ann Intern Med 132 (2), 105–11.

Shackelton, L., Parrish, C., Truyen, U., Holmes, E., Jan. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci U S A 102 (2), 379–384.

Shapiro, B., Rambaut, A., Drummond, A. J., Jan 2006a. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23 (1), 7–9.

Shapiro, B., Rambaut, A., Pybus, O. G., Holmes, E. C., Sep 2006b. A phylogenetic method for detecting positive epistasis in gene sequences and its application to rna virus evolution. Mol Biol Evol 23 (9), 1724–30.

Sharp, P. M., Feb 2002. Origins of human virus diversity. Cell 108 (3), 305–12.

Shi, Z., Hu, Z., Apr 2008. A review of studies on animal reservoirs of the sars coronavirus. Virus Res 133 (1), 74–87.

Simmonds, P., Apr 2001. The origin and evolution of hepatitis viruses in humans. J Gen Virol 82 (Pt 4), 693–712.

Simmonds, P., Nov 2004. Genetic diversity and evolution of hepatitis c virus–15 years on. J Gen Virol 85 (Pt 11), 3173–88.

Simmonds, P., Smith, D., Jul. 1999. Structural constraints on rna virus evolution. J Virol 73 (7), 5787–5794.

Simmonds, P., Tuplin, A., Evans, D., Sep. 2004. Detection of genome-scale ordered rna structure (gors) in genomes of positive-stranded rna viruses: Implications for virus evolution and host persistence. Rna-A Publication of the Rna Society 10 (9), 1337–1351.

Simmons, C. P., Popper, S., Dolocek, C., Chau, T. N. B., Griffiths, M., Dung, N. T. P., Long, T. H., Hoang, D. M., Chau, N. V., Thao, L. T. T., Hien, T. T., Relman, D. A., Farrar, J., Apr 2007. Patterns of host genome-wide gene transcript abundance

in the peripheral blood of patients with acute dengue hemorrhagic fever. J Infect Dis 195 (8), 1097–107.

Smith, D. B., Pathirana, S., Davidson, F., Lawlor, E., Power, J., Yap, P. L., Simmonds, P., Feb 1997. The origin of hepatitis c virus genotypes. J Gen Virol 78 (Pt 2), 321–8.

Smith, G. J. D., Bahl, J., Vijaykrishna, D., Zhang, J., Poon, L. L. M., Chen, H., Webster, R. G., Peiris, J. S. M., Guan, Y., Jul 2009a. Dating the emergence of pandemic influenza viruses. Proc Natl Acad Sci U S A 106 (28), 11709–12.

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., Rambaut, A., Jun 2009b. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. Nature 459 (7250), 1122–5.

Song, H.-D., Tu, C.-C., Zhang, G.-W., Wang, S.-Y., Zheng, K., Lei, L.-C., Chen, Q.-X., Gao, Y.-W., Zhou, H.-Q., Xiang, H., Zheng, H.-J., Chern, S.-W. W., Cheng, F., Pan, C.-M., Xuan, H., Chen, S.-J., Luo, H.-M., Zhou, D.-H., Liu, Y.-F., He, J.-F., Qin, P.-Z., Li, L.-H., Ren, Y.-Q., Liang, W.-J., Yu, Y.-D., Anderson, L., Wang, M., Xu, R.-H., Wu, X.-W., Zheng, H.-Y., Chen, J.-D., Liang, G., Gao, Y., Liao, M., Fang, L., Jiang, L.-Y., Li, H., Chen, F., Di, B., He, L.-J., Lin, J.-Y., Tong, S., Kong, X., Du, L., Hao, P., Tang, H., Bernini, A., Yu, X.-J., Spiga, O., Guo, Z.-M., Pan, H.-Y., He, W.-Z., Manuguerra, J.-C., Fontanet, A., Danchin, A., Niccolai, N., Li, Y.-X., Wu, C.-I., Zhao, G.-P., Feb 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci U S A 102 (7), 2430–5.

Stamatakis, A., Ludwig, T., Meier, H., Feb 2005. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21 (4), 456–63.

Starr, D., 1999. Blood: an epic history of medecine and commerce. Little, Brown & Company, London.

Strimmer, K., Pybus, O. G., Dec 2001. Exploring the demographic history of dna sequences using the generalized skyline plot. Mol Biol Evol 18 (12), 2298–305.

Suchard, M., Kitchen, C., Sinsheimer, J., Weiss, R., Oct. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Systematic Biology 52 (5), 649–664.

Suchard, M., Weiss, R., Sinsheimer, J., Jun. 2001. Bayesian selection of continuous-time markov chain evolutionary models. Mol Biol Evol 18 (6), 1001–1013.

Sullivan, J., Joyce, P., 2005. Model selection in phylogenetics. Annu Rev Ecol Syst 36, 445–466.

Suzuki, Y., Nei, M., Apr 2002. Origin and evolution of influenza virus hemagglutinin genes. Mol Biol Evol 19 (4), 501–9.

Swofford, D. L., 2003. Paup*. phylogenetic analysis using parsimony (*and other methods).

Talbi, C., Lemey, P., Suchard, M. A., Abdelatif, E., Elharrak, M., Nourlil, J., Jalal, N., Faouzi, A., Echevarría, J. E., Vazquez Morón, S., Rambaut, A., Campiz, N., Tatem, A. J., Holmes, E. C., Bourhy, H., 2010. Phylodynamics and human-mediated dispersal of a zoonotic virus. PLoS Pathog 6 (10), e1001166.

Tanaka, Y., Agha, S., Saudy, N., Kurbanov, F., Orito, E., Kato, T., Abo-Zeid, M., Khalaf, M., Miyakawa, Y., Mizokami, M., Feb 2004. Exponential spread of hepatitis c virus genotype 4a in egypt. J Mol Evol 58 (2), 191–5.

Tanaka, Y., Hanada, K., Orito, E., Akahane, Y., Chayama, K., Yoshizawa, H., Sata, M., Ohta, N., Miyakawa, Y., Gojobori, T., Mizokami, M., Jan 2005. Molecular evolutionary analyses implicate injection treatment for schistosomiasis in the initial hepatitis c epidemics in japan. J Hepatol 42 (1), 47–53.

Tang, X. C., Zhang, J. X., Zhang, S. Y., Wang, P., Fan, X. H., Li, L. F., Li, G., Dong, B. Q., Liu, W., Cheung, C. L., Xu, K. M., Song, W. J., Vijaykrishna, D., Poon, L. L. M., Peiris, J. S. M., Smith, G. J. D., Chen, H., Guan, Y., Aug 2006. Prevalence and genetic diversity of coronaviruses in bats from china. J Virol 80 (15), 7481–90.

Tee, K. K., Pybus, O. G., Parker, J., Ng, K. P., Kamarulzaman, A., Takebe, Y., Apr 2009. Estimating the date of origin of an hiv-1 circulating recombinant form. Virology 387 (1), 229–34.

Thorne, J. L., Kishino, H., Painter, I. S., Dec 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15 (12), 1647–57.

Tong, S., Conrardy, C., Ruone, S., Kuzmin, I. V., Guo, X., Tao, Y., Niezgoda, M., Haynes, L., Agwanda, B., Breiman, R. F., Anderson, L. J., Rupprecht, C. E., Mar 2009. Detection of novel sars-like and other coronaviruses in bats from kenya. Emerg Infect Dis 15 (3), 482–5.

Tuplin, A., Wood, J., Evans, D. J., Patel, A. H., Simmonds, P., Jun 2002. Thermodynamic and phylogenetic prediction of rna secondary structures in the coding region of hepatitis c virus. RNA 8 (6), 824–41.

Twiddy, S. S., Holmes, E. C., Rambaut, A., Jan 2003. Inferring the rate and time-scale of dengue virus evolution. Mol Biol Evol 20 (1), 122–9.

van de Laar, T. J. W., Molenkamp, R., van den Berg, C., Schinkel, J., Beld, M. G. H. M., Prins, M., Coutinho, R. A., Bruisten, S. M., Oct 2009. Frequent hcv reinfection and superinfection in a cohort of injecting drug users in amsterdam. J Hepatol 51 (4), 667–74.

van der Poel, C. L., 1999. Hepatitis c virus and blood transfusion: past and present risks. J Hepatol 31 Suppl 1, 101–6.

Vasilakis, N., Cardosa, J., Hanley, K. A., Holmes, E. C., Weaver, S. C., Jul 2011. Fever from the forest: prospects for the continued emergence of sylvatic dengue virus and its impact on public health. Nat Rev Microbiol 9 (7), 532–41.

Vijaykrishna, D., Smith, G. J. D., Pybus, O. G., Zhu, H., Bhatt, S., Poon, L. L. M., Riley, S., Bahl, J., Ma, S. K., Cheung, C. L., Perera, R. A. P. M., Chen, H., Shortridge, K. F., Webby, R. J., Webster, R. G., Guan, Y., Peiris, J. S. M., May 2011. Long-term evolution and transmission dynamics of swine influenza a virus. Nature 473 (7348), 519–22.

Vijaykrishna, D., Smith, G. J. D., Zhang, J. X., Peiris, J. S. M., Chen, H., Guan, Y., Apr 2007. Evolutionary insights into the ecology of coronaviruses. J Virol 81 (8), 4012–20.

Villarreal, L. P., Defilippis, V. R., Gottlieb, K. A., Jun 2000. Acute and persistent viral life strategies and their relationship to emerging diseases. Virology 272 (1), 1–6.

Wang, M., Yan, M., Xu, H., Liang, W., Kan, B., Zheng, B., Chen, H., Zheng, H., Xu, Y., Zhang, E., Wang, H., Ye, J., Li, G., Li, M., Cui, Z., Liu, Y.-F., Guo, R.-T., Liu, X.-N., Zhan, L.-H., Zhou, D.-H., Zhao, A., Hai, R., Yu, D., Guan, Y., Xu, J., Dec 2005. Sars-cov infection in a restaurant from palm civet. Emerg Infect Dis 11 (12), 1860–5.

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., Kawaoka, Y., Mar 1992. Evolution and ecology of influenza a viruses. Microbiol Rev 56 (1), 152–79.

Wertheim, J. O., 2010. The re-emergence of h1n1 influenza virus in 1977: a cautionary tale for estimating divergence times using biologically unrealistic sampling dates. PLoS One 5 (6), e11184.

Wertheim, J. O., Kosakovsky Pond, S. L., Jun 2011. Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol.

Whelan, S., Goldman, N., May 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18 (5), 691–9.

WHO, December 2008. Who factsheet: Ebola hemarrhagic fever.
URL `http://www.who.int/mediacentre/factsheets/fs103/en/`

Wiuf, C., Christensen, T., Hein, J., Oct 2001. A simulation study of the reliability of recombination detection methods. Mol Biol Evol 18 (10), 1929–39.

Wolfe, N. D., Daszak, P., Kilpatrick, A. M., Burke, D. S., Dec 2005. Bushmeat hunting, deforestation, and prediction of zoonoses emergence. Emerg Infect Dis 11 (12), 1822–7.

Wolfe, N. D., Dunavan, C. P., Diamond, J., May 2007. Origins of major human infectious diseases. Nature 447 (7142), 279–83.

Woo, P. C. Y., Lau, S. K. P., Li, K. S. M., Poon, R. W. S., Wong, B. H. L., Tsoi, H.-w., Yip, B. C. K., Huang, Y., Chan, K.-h., Yuen, K.-y., Jul 2006. Molecular diversity of coronaviruses in bats. Virology 351 (1), 180–7.

Woolhouse, M. E. J., 2002. Population biology of emerging and re-emerging pathogens. Trends Microbiol 10 (10 Suppl), S3–7.

Woolhouse, M. E. J., Haydon, D. T., Antia, R., May 2005. Emerging pathogens: the epidemiology and evolution of species jumps. Trends Ecol Evol 20 (5), 238–44.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T. P., Wolinsky, S. M., Oct 2008. Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. Nature 455 (7213), 661–4.

Worobey, M., Holmes, E. C., Oct 1999. Evolutionary aspects of recombination in rna viruses. J Gen Virol 80 (Pt 10), 2535–43.

Xu, X., Subbarao, K., Cox, N., Guo, Y., Aug. 1999. Genetic characterization of the pathogenic influenza a/goose/guangdong/1/96 (h5n1) virus: Similarity of its hemagglutinin gene to those of h5n1 viruses from the 1997 outbreaks in hong kong. Virology 261 (1), 15–19.

Yang, Z., Rannala, B., Jul 1997. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. Mol Biol Evol 14 (7), 717–24.

Yip, C. W., Hon, C. C., Shi, M., Lam, T. T.-Y., Chow, K. Y.-C., Zeng, F., Leung, F. C.-C., Dec 2009. Phylogenetic perspectives on the epidemiology and origins of sars and sars-like coronaviruses. Infect Genet Evol 9 (6), 1185–96.

Yoder, A. D., Yang, Z., Jul 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17 (7), 1081–90.

Yuan, J., Hon, C.-C., Li, Y., Wang, D., Xu, G., Zhang, H., Zhou, P., Poon, L. L. M., Lam, T. T.-Y., Leung, F. C.-C., Shi, Z., Apr 2010. Intraspecies diversity of sars-like coronaviruses in rhinolophus sinicus and its implications for the origin of sars coronaviruses in humans. J Gen Virol 91 (Pt 4), 1058–62.

Yurovsky, A., Moret, B. M. E., 2011. Fluref, an automated flu virus reassortment finder based on phylogenetic trees. BMC Genomics 12, doi:10.1186/1471–2164–12–S2–S3.

Zhai, W., Slatkin, M., Nielsen, R., Sep 2007. Exploring variation in the d(n)/d(s) ratio among sites and lineages using mutational mappings: applications to the influenza virus. J Mol Evol 65 (3), 340–8.

Zhang, C., Mammen, Jr, M. P., Chinnawirotpisan, P., Klungthong, C., Rodpradit, P., Monkongdee, P., Nimmannitya, S., Kalayanarooj, S., Holmes, E. C., Dec 2005. Clade replacements in dengue virus serotypes 1 and 3 are associated with changing serotype prevalence. J Virol 79 (24), 15123–30.

Zhong, N. S., Zheng, B. J., Li, Y. M., Poon, Xie, Z. H., Chan, K. H., Li, P. H., Tan, S. Y., Chang, Q., Xie, J. P., Liu, X. Q., Xu, J., Li, D. X., Yuen, K. Y., Peiris, Guan, Y., Oct 2003. Epidemiology and cause of severe acute respiratory syndrome (sars) in guangdong, people's republic of china, in february, 2003. Lancet 362 (9393), 1353–8.

A

Supplementary Material: Chapter 3

Table A.1: List of SARS-CoV sequences and their corresponding isolation dates used in the study.

Accession Number	**Isolate Name**	**Date of Isolation**	**Host Species**
DQ648856	Bat CoV 273/2005	7-Nov-2004	Bat
DQ648857	Bat CoV 279/2005	8-Nov-2004	Bat
DQ084200	Bat SARS-CoV HKU3-3	17-Mar-2005	Bat
DQ084199	Bat SARS-CoV HKU3-2	24-Feb-2005	Bat
DQ022305.2	Bat SARS-CoV HKU3-1	17-Feb-2005	Bat
DQ412042	Bat SARS-CoV Rf1/2004	7-Nov-2004	Bat
DQ412043	Bat SARS-CoV Rm1/2004	8-Nov-2004	Bat
DQ071615	Bat SARS-CoV RP3/2004	2-Dec-2004	Bat
GQ153539	Bat SARS-CoV HKU3-4	20-Jul-2005	Bat
GQ153540	Bat SARS-CoV HKU3-5	20-Sep-05	Bat
GQ153541	Bat SARS-CoV HKU3-6	16-Dec-05	Bat
GQ153542	Bat SARS-CoV HKU3-7	15-Feb-06	Bat
GQ153543	Bat SARS-CoV HKU3-8	15-Feb-06	Bat
GQ153544	Bat SARS-CoV HKU3-9	28-Aug-06	Bat
GQ153545	Bat SARS-CoV HKU3-10	28-Aug-06	Bat
GQ153546	Bat SARS-CoV HKU3-11	07-Mar-07	Bat
GQ153547	Bat SARS-CoV HKU3-12	15-May-07	Bat
GQ153548	Bat SARS-CoV HKU3-13	15-Nov-07	Bat
FJ588686	Bat SARS-CoV Rs672/2006	13-Nov-06	Bat
AY304486.1	SARS-CoV SZ3	7-May-2003	Civet
AY304487.1	SARS-CoV SZ13	7-May-2003	Civet
AY304488.1	SARS-CoV SZ16	7-May-2003	Civet
AY304489.1	SARS-CoV SZ1	7-May-2003	Civet
AY304490.1	SARS-CoV GZ43	20-Feb-2003	Human
AY304491.1	SARS-CoV GZ60	21-Feb-2003	Human
AY304495.1	SARS-CoV GZ50	20-Feb-2003	Human
AY515512.1	SARS-CoV HC/SZ/61/03	20-Dec-2003	Human

Continued on Next Page...

Table A.1 – Continued

Accession Number	Isolate Name	Date of Isolation	Host Species
AY545914.1	SARS-CoV HC/SZ/79/03	29-Dec-2003	Human
AY545915.1	SARS-CoV HC/SZ/DM1/03	6-Nov-2003	Human
AY545916.1	SARS-CoV HC/SZ/266/03	22-Oct-2003	Human
AY545917.1	SARS-CoV HC/GZ/81/03	31-Dec-2003	Human
AY545918.1	SARS-CoV HC/GZ/32/03	12-Dec-2003	Human
AY545919.1	SARS-CoV CFB/SZ/DM94/03	29-Dec-2003	Human
AY274119	SARS-CoV TOR2	13-Apr-03	Human
AY278487	SARS-CoV BJ02	21-Apr-03	Human
AY278488	SARS-CoV BJ01	7-Mar-03	Human
AY278489	SARS-CoV GD01	13-Apr-03	Human
AY278490	SARS-CoV BJ03	3-Jun-03	Human
AY279354	SARS-CoV BJ04	3-Mar-03	Human
AY282752	SARS-CoV CUHKSu10	13-Mar-03	Human
AY283796	SARS-CoV Sin2679	15-Mar-03	Human
AY291315	SARS-CoV Frankfurt1	15-Mar-03	Human
AY297028	SARS-CoV ZJ01	12-May-03	Human
AY304493	SARS-CoV HKU66078	24-Mar-03	Human
AY313906	SARS-CoV GD69	3-Aug-03	Human
AY390556	SARS-CoV GZ02	11-Feb-03	Human
AY394850	SARS-CoV WHU	3-May-03	Human
AY394977	SARS-CoV GZ-A	22-Feb-03	Human
AY394978	SARS-CoV GZ-B	15-Mar-03	Human
AY394999	SARS-CoV LC2	15-May-03	Human
AY502924	SARS-CoV TW11	15-May-03	Human
AY559084	SARS-CoV Sin3765V	15-Apr-03	Human
AY559087	SARS-CoV Sin3725V	15-Apr-03	Human
AY559093	SARS-CoV Sin845	7-Apr-03	Human
AY568539	SARS-CoV GZ0301	22-Dec-03	Human
AY572034	SARS-CoV Cv007G	5-Jan-04	Human
AY572038	SARS-CoV Cv020G	2-Jan-04	Human
AY613948	SARS-CoV PC413	10-Jan-04	Human
AY613949	SARS-CoV PC4136	5-Jan-04	Human
AY613950	SARS-CoV PC4227	5-Jan-04	Human
AY686863	SARS-CoV A022G	5-Jan-04	Human
AY686864	SARS-CoV B039G	5-Jan-04	Human

B

Supplementary Material: Chapter 4

Methods

Sequencing of new HCV 2k/1b isolates from Amsterdam

The HCV RNA levels of the patients were determined using the Siemens Versant HCV RNA 3.0 assay (bDNA) with a linear dynamic range of 615 to 7.7x106 IU/mL and a LOD of 615 IU/mL. HCV RNA was isolated from 200l plasma using the purification method described by Boom et al. (1990). cDNA was generated using random hexamer primers as described before (Boom et al., 1990).

Amplification of a 265nt fragment of the 5UTR fragment was performed in a 25 L volume using HCV47F: GTGAGGAACTACTGTCTTCACG as the forward primer and HCV312R: ACTCGCAAGCACCCTATCAGG as the reverse primer and using FastStart Taq DNA Polymerase with additional ready-to-use PCR Grade PCR Mix kit (ROCHE Diagnostics GmbH). Final concentrations of 0.25 M/L of primer, 0.5 mM/L deoxynucleoside triphosphate, 0.1 g/L bovine serum albumin, and 2.5 mM MgCl2 were used.

The amplification was performed using a conventional-PCR with the following cycling conditions: 2 min at 50 C and 10 min at 95C, followed by 45 cycles each consisting of 30 sec at 95C, 30 seconds at 55C, and 1 min at 72C. Amplicons were purified from a 1% agarose gel as described by Boom et al. (1990). Amplification of a 724 nt fragment of the NS5B region was performed as previously described (Murphy et al., 2007). Amplification of E1 was performed in a 25 L volume using HCV1b/2 F: GCGTGAGRGTCCTGGAG as the forward primer and HCV1b/2 R: TGCCARCARTANGGCYTCAT as the reverse primer, using the same amplification conditions as mentioned above. Amplicons were also purified from a 1% agarose gel as described by (Boom et al., 1990). Sequencing was performed using HCV1b/2F seq: CTTCYTACTAGCTCTYTTGTCTT as the forward sequencing primer and HCV1b/2R seq: TGCCAACTGCCRTTGGTGT as the reverse sequencing primer.

To confirm the viral variants were recombinants, a 234 nt fragment harbouring the known break point in NS2 was also amplified and sequenced, Amplification and sequencing of the NS2 breakpoint was performed using the HCV2K_F: GCACGCCATACTTCGTCAGAG as the forward primer and HCV1B_R: CAGGTAATGATCTTGGTCTCCATGT as the reverse primer also using the same cycling conditions as mentioned above.

Tables

Table B.1: List of HCV isolates used in the study and the corresponding accession numbers in GenBank

Isolate Name	**NS5B**	**Core/E1**
1b_TJ_TAJ101_2006	AB330319	AB330348
1b_TJ_TAJ81_2006	AB330340	AB330362
1b_US_V121_1992	EU155324	EU155324
1b_US_V127_1992	EU155328	EU155328
1b_US_V128_1992	EU155329	EU155329
1b_US_V131_1990	EU155331	EU155331
1b_US_V135_1992	EU234062	EU234062
1b_US_V138_1989	EU482849	EU482849
1b_US_V140_1991	EU255962	EU255962
1b_US_V141_1990	EU155337	EU155337
1b_US_V144_2002	EU256001	EU256001
1b_US_V154_2001	EU660388	EU660388
1b_US_V157_2003	EU155227	EU155227
1b_US_V159_2004	EU155229	EU155229
1b_US_V164_2002	EU155232	EU155232
1b_US_V1711_2007	FJ024277	FJ024277
1b_US_V1715_2007	FJ024279	FJ024279
1b_US_V1748_2007	FJ390396	FJ390396
1b_US_V1750_2008	FJ390398	FJ390398
1B_US_217302382_BID_V2148_1999	FJ478453	FJ478453
1b_CH_V272_2003	EU482859	EU482859
1b_CH_V275_2003	EU155357	EU155357
1b_CH_V279_2004	EU256075	EU256075
1b_CH_V286_2005	EU256079	EU256079
1B_SWISS_190684402_BID_V292_2002	EU256080	EU256080
1b_CH_V296_2001	EU155368	EU155368
1b_CH_V297_2002	EU155369	EU155369
1b_CH_V298_2005	EU256082	EU256082
1b_CH_V299_2005	EU256083	EU256083
1b_CH_V302_2003	EU482860	EU482860
1B_SWISS_190684410_BID_V303	EU256084	EU256084
1b_CH_V304_2004	EU862837	EU862837
1b_CH_V306_2004	EU155372	EU155372
1b_CH_V307_2005	EU155373	EU155373
1b_CH_V309_2006	EU155375	EU155375
1b_CH_V311_2006	EU155376	EU155376
1b_CH_V313_2006	EU155377	EU155377
1b_US_V341_2003	EU155300	EU155300

Continued on Next Page...

Table B.1 – Continued

Isolate Name	NS5B	Core/E1
1b_US_V344_2001	EU155301	EU155301
1b_US_V346_2001	EU155303	EU155303
1b_US_V347_2003	EU155304	EU155304
1B_US_190171145_BID_V348	EU256099	EU256099
1b_US_V352_2002	EU155306	EU155306
1b_US_V353_2002	EU155307	EU155307
1b_US_V355_2006	EU660386	EU660386
1b_US_V363_2006	EU155253	EU155253
1b_US_V364_2006	EU256061	EU256061
1b_US_V366_2006	EU155255	EU155255
1b_US_V369_2006	EU155257	EU155257
1b_US_V374_2006	EU155259	EU155259
1b_US_V379_2005	EU155261	EU155261
1b_US_V382_2003	EU256066	EU256066
1b_US_V384_2005	EU155263	EU155263
1b_US_V416_2001	EU155315	EU155315
1b_US_V420_2002	EU155317	EU155317
1b_US_V443_2001	EU256054	EU256054
1b_US_V458_2006	EU482886	EU482886
1b_US_V463_2006	EU256092	EU256092
1b_DE_V502_2004	EU155381	EU155381
1B_AU_HCV_A_AJ000009	AJ000009	AJ000009
1B_CHIN_HCV_S_AY460204	AY460204	AY460204
1B_DE_4753720_AJ132997	AJ132997	AJ132997
1B_DE_5748510_AJ238800	AJ238800	AJ238800
1B_IRE_56342240_AB154205	AB154205	AB154205
1B_JP_HCVT221_AB049101	AB049101	AB049101
1B_JP_HCV_JS_D85516	D85516	D85516
1B_JP_HCV_J_D90208	D90208	D90208
1B_JP_MD1_911_AF165046	AF165046	AF165046
1B_RU_N589_AY587844	AY587844	AY587844
1B_TR_HCV_TR1_AF483269	AF483269	AF483269
1B_TW_HCU89019_U89019	U89019	U89019
1b2k_AZ_01AZ051_2000	FJ435529	FJ435462
1b2k_AZ_01AZ082_2000	FJ435544	FJ435480
1b2k_AZ_02AZ105_2001	FJ435550	FJ435490
1b2k_AZ_02AZ114_2001	FJ435556	FJ435497
1b2k_AZ_02AZ129_2001	FJ435564	FJ435505
1b2k_AZ_02AZ139_2001	FJ435572	FJ435514
1b2k_CY_CYHCV037_2005	EU684614	EU684686
1b2k_CY_CYHCV093_2007	EU684649	EU684728
c_1b2k_AM_P077_2006	JF949902	JF949908

Continued on Next Page...

Table B.1 – Continued

Isolate Name	**NS5B**	**Core/E1**
c_1b2k_AM_P079_2006	JF949897	JF949903
c_1b2k_AM_P108_2007	JF949898	JF949904
c_1b2k_AM_P135_2005	JF949899	JF949905
c_1b2k_AM_P159_2007	JF949900	JF949906
c_1b2k_AM_P179_2000	JF949901	JF949907
c_1b2k_FR_M21_2007	FJ821465	FJ821465
c_1b2k_IE_HC9A99966_2006	AB327058	AB327018
c_1b2k_RU_747_1999	AF388411	AY070214
c_1b2k_RU_796_1999	AF388412	AY070215
c_1b2k_RU_ALT30_2000	AB327055	AB327015
c_1b2k_RU_HIA1002_2003	DQ001221	AB327011
c_1b2k_RU_KNG318_2002	AY764172	AB327010
c_1b2k_RU_KNG327_2002	AY764176	AB327012
c_1b2k_RU_N687_1999	AY587845	AY587845
c_1b2k_RU_PSA108_2005	AB327053	AB327013
c_1b2k_RU_PSA62_2005	AB327054	AB327014
c_1b2k_UZ_AZ15	AB327056	AB327016
c_1b2k_UZ_UZIDU19_2006	AB327120	AB327122
2K_MOLD_VAT96_AB031663	AB031663	AB031663
2k_AZ_02AZ149_2001	FJ435578	FJ435522
2k_FR_G2MP003	DQ220879	AB327022
2k_FR_G2MP004	DQ220876	AB327023
2k_FR_G2MP014	DQ220877	AB327025
2k_FR_G2MP022	DQ220882	AB327026
2k_FR_G2MP118	DQ220910	AB327029
2k_FR_G2MP120	DQ220893	AB327031
2k_FR_G2MP121	DQ220896	AB327032
2k_FR_G2MP124	DQ220905	AB327035
2k_FR_G2MP125	DQ220894	AB327036
2k_FR_G2MP126	DQ220900	AB327037
2k_FR_G2MP127	DQ220895	AB327038
2k_FR_G2MP128	DQ220889	AB327039
2k_FR_G2MP135	DQ220891	AB327043

C

Supplementary Material: Chapter 6

Methods

Patient population

The dengue patients from whom DENV whole genome sequences were determined were enrolled in one of two prospective studies at the Hospital for Tropical Diseases in Ho Chi Minh City, Viet Nam or at Dong Thap Hospital, Dong Thap Province, Viet Nam. The median age of these patients was 12 years (interquartile range 7-17 years) and 51% were male. Serological investigations (IgM and IgG capture ELISAs) were performed using paired plasma samples using methods described previously (Hang et al., 2010). DENV serotype and viraemia levels were determined using an internally-controlled real-time RT-PCR assay that has been described previously (Simmons et al., 2007).

Genomic sequencing

Viral genomes were sequenced using the Broad Institutes capillary sequencing (Applied Biosystems) directed amplification viral sequencing pipeline http://www.broadinstitute.org /scientific-community/science/projects/viral-genomics-initiative). This sequencing effort was part of the Broad Institutes Genome Resources in Dengue Consortium (GRID) project. Viral RNA was isolated from diagnostic plasma samples (QIAmp viral RNA mini kit, Qiagen) and the RNA genome reverse transcribed to cDNA with superscript III reverse transcriptase (Invitrogen), random hexamers (Roche) and a specific oligonucleotide targeting the 3 end of the target genome sequences (nt 10868 to 10890, AGAACCTGTTGATTCAACAGCAC). cDNA was then amplified using a high fidelity DNA polymerase (pfu Ultra II, Stratagene) and a pool of specific primers to produce 14 overlapping amplicons of 1.5 to 2 kb in size for a physical coverage of 2-fold across the target genome (nt 40 to 10649). Amplicons were then sequenced in the forward and reverse direction using primer panels consisting of 96 specific primer pairs, tailed with M13 forward and reverse primer sequence, producing 500700 bp amplicons from the target viral genome. Amplicons were then sequenced in the forward and reverse direction using M13 primer. Total coverage delivered post amplification and sequencing was 8-fold. Resulting sequence reads were assembled de novo using the Broad Institutes AV454 assembly algorithm (Henn *et al.* 2011. in review) and a reference-based annotation algorithm. All whole genome sequences newly determined here have been deposited in GenBank and assigned accession numbers (see Table S1 in Raghwani et al. (2011)).

Ethics Statement

Patients (or their parents/guardians) gave written informed consent to participate in each of the studies. The study protocols were approved by the Hospital for Tropical Diseases and the Oxford University Tropical Research Ethical Committee.

D

Related Publications

nature Vol 459|25 June 2009|doi:10.1038/nature08182

LETTERS

Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic

Gavin J. D. Smith[1], Dhanasekaran Vijaykrishna[1], Justin Bahl[1], Samantha J. Lycett[2], Michael Worobey[3], Oliver G. Pybus[4], Siu Kit Ma[1], Chung Lam Cheung[1], Jayna Raghwani[2], Samir Bhatt[4], J. S. Malik Peiris[1], Yi Guan[1] & Andrew Rambaut[2]

In March and early April 2009, a new swine-origin influenza A (H1N1) virus (S-OIV) emerged in Mexico and the United States[1]. During the first few weeks of surveillance, the virus spread worldwide to 30 countries (as of May 11) by human-to-human transmission, causing the World Health Organization to raise its pandemic alert to level 5 of 6. This virus has the potential to develop into the first influenza pandemic of the twenty-first century. Here we use evolutionary analysis to estimate the time-scale of the origins and the early development of the S-OIV epidemic. We show that it was derived from several viruses circulating in swine, and that the initial transmission to humans occurred several months before recognition of the outbreak. A phylogenetic estimate of the gaps in genetic surveillance indicates a long period of unsampled ancestry before the S-OIV outbreak, suggesting that the reassortment of swine lineages may have occurred years before emergence in humans, and that the multiple genetic ancestry of S-OIV is not indicative of an artificial origin. Furthermore, the unsampled history of the epidemic means that the nature and location of the genetically closest swine viruses reveal little about the immediate origin of the epidemic, despite the fact that we included a panel of closely related and previously unpublished swine influenza isolates. Our results highlight the need for systematic surveillance of influenza in swine, and provide evidence that the mixing of new genetic elements in swine can result in the emergence of viruses with pandemic potential in humans[2].

Initial genetic characterization of the S-OIV outbreak by the United States Centers for Disease Control suggested swine as its probable source, on the basis of sequence similarity to previously reported swine influenza isolates[1]. Classical swine H1N1 viruses have circulated in pigs in North America and other regions for at least 80 years[3]. In 1998, a new triple-reassortant H3N2 virus—comprising genes from classical swine H1N1, North American avian, and human H3N2 (A/Sydney/5/97-like) influenza—was reported as the cause of outbreaks in North American swine, with subsequent establishment in pig populations[4,5]. Co-circulation and mixing of the triple-reassortant H3N2 with established swine lineages subsequently generated further H1N1 and H1N2 reassortant swine viruses[6–8], which have caused sporadic human infections in the United States since 2005 (refs 6, 7). Consequently, human infection with H1N1 swine influenza has been a nationally notifiable disease in the United States since 2007 (ref. 9). In Europe, an avian H1N1 virus was introduced to pigs ('avian-like' swine H1N1) and first detected in Belgium in 1979 (ref. 10). This lineage became established and gradually replaced classical swine H1N1 viruses, and also reassorted in pigs with human H3N2 viruses (A/Port Chalmers/1/1973-like)[11]. It is noteworthy that, until now, there has been no evidence of Eurasian avian-like swine H1N1 circulating in North American pigs. In Asia, the classical swine influenza lineage circulates, in addition to other identified viruses, including human H3N2, Eurasian avian-like H1N1, and North American triple-reassortant H3N2 (refs 12, 13).

Using comprehensive phylogenetic analyses, we have estimated a temporal reconstruction of the complex reassortment history of the S-OIV outbreak, summarized in Fig. 1 (Methods). Our analyses showed that each segment of the S-OIV genome was nested within a well-established swine influenza lineage (that is, a lineage circulating primarily in swine for >10 years before the current outbreak). The most parsimonious interpretation of these results is therefore that the progenitor of the S-OIV epidemic originated in pigs. Some transmission of swine influenza has, however, been observed in secondary hosts in North America, for example, in turkeys[14]. Although the precise evolutionary pathway of the genesis of S-OIV is greatly hindered by the lack of surveillance data (see later), we can conclude that the polymerase genes, plus HA, NP and NS, emerged from a triple-reassortant virus circulating in North American swine. The source triple-reassortant itself comprised genes derived from avian (PB2 and PA), human H3N2 (PB1) and classical swine (HA, NP and NS) lineages. In contrast, the NA and M gene segments have their origin in the Eurasian avian-like swine H1N1 lineage. Phylogenetic analyses from the early days of the outbreak, on the basis of the first publicly available sequences, quickly established this multiple genetic origin (refs 8, 15, 16 and http://influenza.bio.ed.ac.uk).

Given that S-OIV contains genes of Eurasian origin, we included in our phylogenetic analyses 15 newly sequenced swine influenza viruses from Hong Kong, sampled in the course of a surveillance program conducted since the early 1990s. The viruses were a mixture of seven H1N1 and eight H1N2 subtypes, and viruses belonging to the classical, Eurasian avian-like, and triple-reassortant swine lineages were all present. Both Eurasian and triple-reassortant strains were isolated in Hong Kong in 2009. Extensive reassortment among these three virus lineages was also observed from the Hong Kong surveillance data (Supplementary Table 3), with reassortment between Eurasian avian-like and triple-reassortant swine lineages occurring as early as 2003 (for example, Sw/HK/78/2003).

Notably, for the PB1, HA and M genes, some of these newly generated sequences are more similar to the S-OIV epidemic than any previously reported isolates (Supplementary Fig. 2). Notably, seven out of eight genomic segments found in a single 2004 isolate (Sw/HK/915/04 (H1N2)) were located in a sister lineage to the current outbreak. Not only does this suggest that the precursors of S-OIV were swine viruses,

[1]State Key Laboratory of Emerging Infectious Diseases & Department of Microbiology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Pokfulam, Hong Kong SAR, China. [2]Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, King's Buildings, Edinburgh EH9 3JT, UK. [3]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85705, USA. [4]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

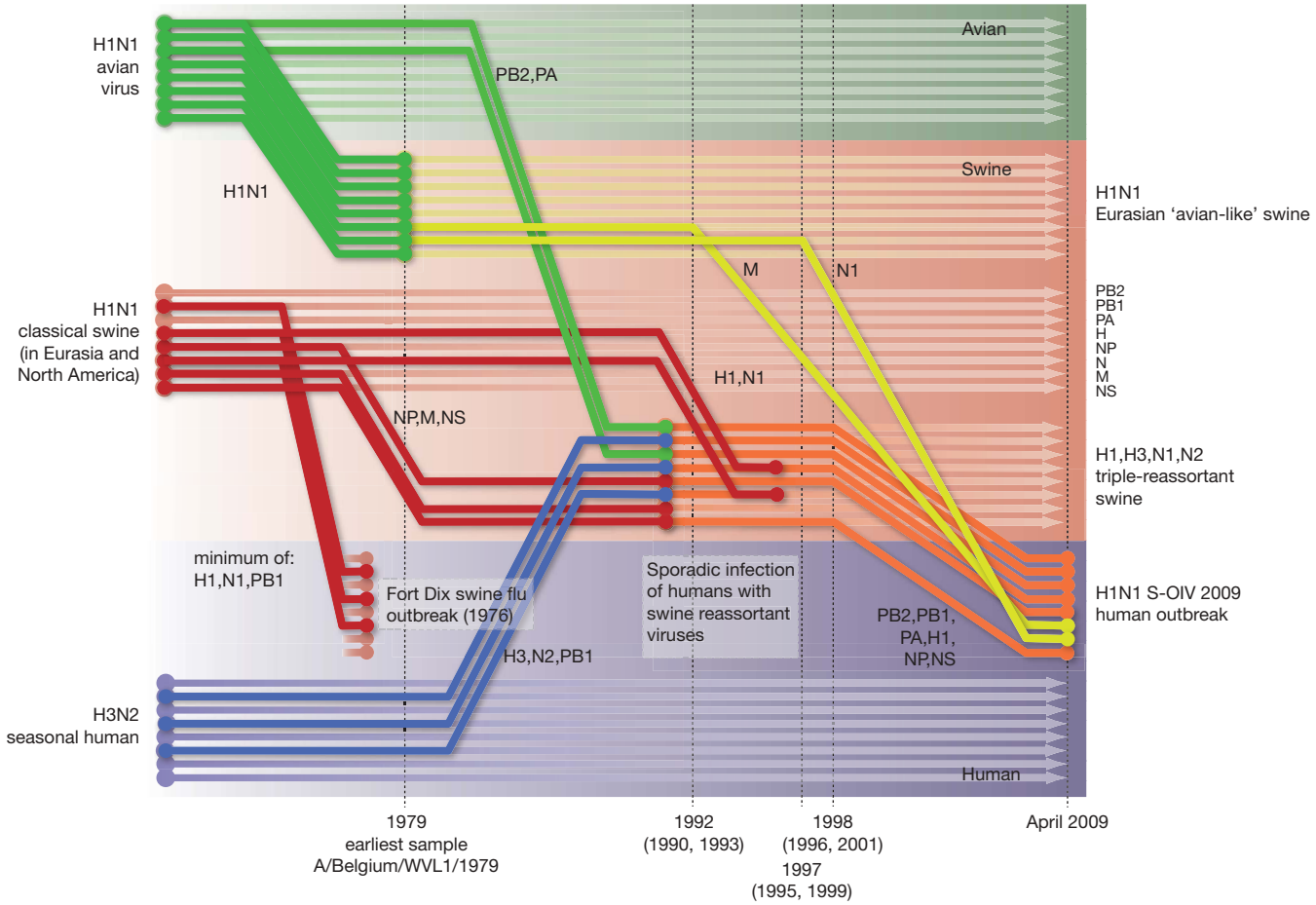


Figure 1 | Reconstruction of the sequence of reassortment events leading up to the emergence of S-OIV. Shaded boxes represent host species; avian (green), swine (red) and human (grey). Coloured lines represent interspecies-transmission pathways of influenza genes. The eight genomic segments are represented as parallel lines in descending order of size. Dates marked with dashed vertical lines on 'elbows' indicate the mean time of divergence of the S-OIV genes from corresponding virus lineages. Reassortment events not involved with the emergence of human disease are omitted. Fort Dix refers to the last major outbreak of S-OIV in humans. The first triple-reassortant swine viruses were detected in 1998, but to improve clarity the origin of this lineage is placed earlier.

but also that they were geographically widely distributed. Crucially, however, the observation of a sister relationship between the current outbreak virus and Sw/HK/915/04 cannot be interpreted as evidence for a Eurasian origin of the outbreak, owing to the long branch of the phylogeny leading to the 2009 human strains (Fig. 2 and Table 1). This branch must represent either an increased rate of evolution leading to the outbreak, or a long period during which the ancestors of the current epidemic went unsampled. To test these hypotheses, we regressed genetic divergence against sampling date for each gene, and found in favour of the latter: the evolutionary rate preceding the S-OIV epidemic is entirely typical for swine influenza (Supplementary Figs 2 and 3).

Therefore, to quantify the period of unsampled diversity, and to estimate the date of origin for the S-OIV outbreak, we performed a Bayesian molecular clock analysis for each gene (Methods). We also estimated the rate of evolution and time of the most recent common ancestor (TMRCA) of a set of genome sequences sampled from the S-OIV epidemic (between March and May 2009; isolates listed in Supplementary Table 4). We found that the common ancestor of the S-OIV outbreak and the closest related swine viruses existed between 9.2 and 17.2 years ago, depending on the genomic segment, hence the ancestors of the epidemic have been circulating undetected for about a decade. In contrast, the currently sampled S-OIV shared a common ancestor around January 2009 (no earlier than August 2008; Table 1). The long, unsampled history observed for every segment suggests that the reassortment of Eurasian and North American swine lineages may not have occurred recently, and it is possible that this single reassortant lineage has been cryptically circulating rather than two distinct lineages of swine flu. Thus, this genomic structure may have been circulating in pigs for several years before emergence in humans, and we urge caution in making inferences about human adaptation on the basis of the ancestry of the individual genes.

A search for amino acid residues in the S-OIV outbreak sequences that have been previously identified as phenotypic markers showed no evidence of virulence-associated variation or adaptations to human hosts[17–19], consistent with the outbreak being of swine origin and causing relatively mild symptoms. Full molecular characterization of the human swine H1N1 viruses is provided in Supplementary Information.

We did detect a difference in the viral molecular evolution in the outbreak clade when compared to that observed in related swine influenza sequences: all S-OIV genes showed a comparatively higher non-synonymous to synonymous (d_N/d_S) substitution rate ratio (Supplementary Tables 1 and 2). This d_N/d_S ratio rise could be due to the increased detection of mildly deleterious mutations resulting from intensive epidemic surveillance; such mutations would more typically be eliminated and escape detection[20]. Alternatively, these mutations could be adaptations to the new host species.

Because this d_N/d_S ratio rise may affect our estimate of the TMRCA of the S-OIV outbreak strains (which was estimated using long-term rates of swine influenza evolution), we compared the mean d_N/d_S values of outbreak versus non-outbreak data sets, thereby approximating the degree of excess of non-synonymous mutations in the outbreak sequences (Methods). Once the d_N/d_S ratio rise is corrected

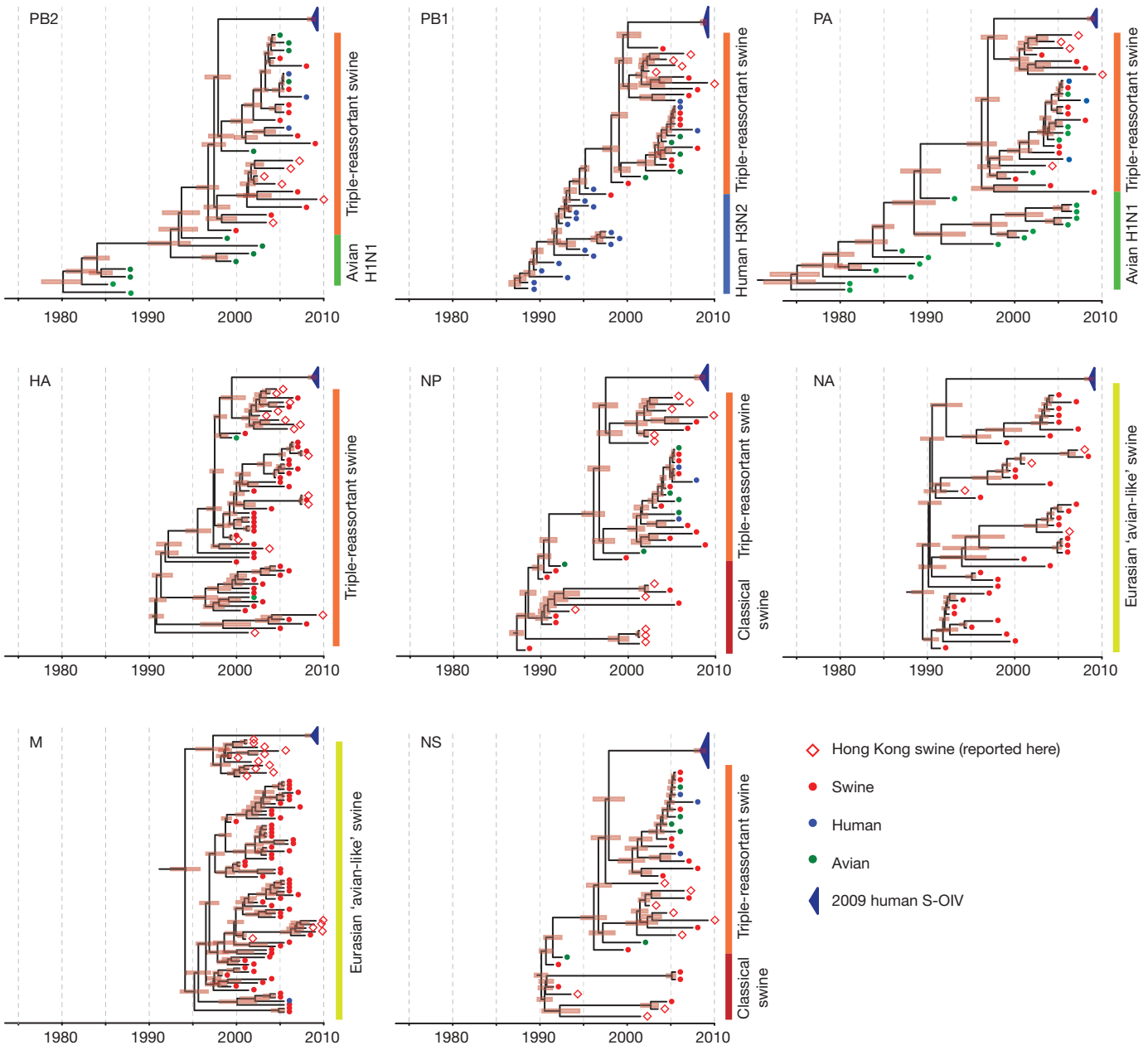


Figure 2 | Genetic relationships and timing of S-OIV for each genomic segment. Symbols represent sampled viruses on a timescale of when they were sampled and coloured by host species (pigs, red; humans, blue; birds, green). Internal nodes are reconstructed common ancestors with 95% credible intervals on their date given by the red bars. The S-OIV outbreak strains are represented by a blue triangle, with the apex representing the common ancestor of these.

for, the mean TMRCA of the S-OIV outbreak became 1 to 5 months more recent for each gene (Supplementary Tables 1 and 2). Furthermore, the adjusted TMRCA estimates are more uniform across genes, and are more similar to that obtained using internally calibrated S-OIV complete genomes (Table 1; a comparable estimate for the TMRCA of the HA gene only was recently reported[21]). Irrespective of whether the d_N/d_S ratio rise is due to increased detection of deleterious mutations or to increased adaptive evolution, its

Table 1 | Time of most recent common ancestors for the S-OIV outbreak

Gene	TMRCA of outbreak samples	Duration of unsampled diversity (years)	Mean evolutionary rate $\times 10^{-3}$ (subst. per site per year)
HA	28 Aug 2008 (1 Apr 2008, 2 Jan 2009)	9.80 (8.41, 11.02)	3.67 (3.41, 3.92)
MP	3 Aug 2008 (8 Dec 2007, 5 Feb 2009)	11.82 (10.17, 13.74)	2.55 (2.19, 2.93)
NA	8 Aug 2008 (23 Feb 2008, 26 Dec 2008)	17.15 (15.40, 18.88)	3.65 (3.22, 4.12)
NP	27 Mar 2008 (15 Sep 2007, 19 Sep 2008)	11.83 (10.53, 13.23)	2.59 (2.34, 2.84)
NS	21 May 2008 (30 Sep 2007, 27 Nov 2008)	11.47 (9.75, 13.21)	2.62 (2.32, 2.92)
PA	7 Oct 2008 (1 Jun 2008, 1 Feb 2009)	11.70 (10.25, 13.10)	2.45 (2.20, 2.69)
PB1	24 Oct 2008 (8 Jul 2008, 25 Jan 2009)	9.24 (7.59, 10.48)	2.34 (2.13, 2.53)
PB2	9 Sep 2008 (12 Apr 2008, 9 Jan 2009)	11.26 (9.93, 12.69)	2.60 (2.29, 2.92)
Genome*	21 Jan 2009 (3 Aug 2008, 13 Mar 2009)	N/A	3.66 (0.61, 6.58)

The values in parentheses represent the 95% credible intervals.
* This data set comprises complete or partial genomes of swine-origin influenza A (H1N1) virus outbreak isolates sampled predominantly in the United States between March and May 2009.

presence may be a general feature of intensively sampled emerging epidemics, and should be accounted for in the evolutionary analysis of such events.

Movement of live pigs between Eurasia and North America seems to have facilitated the mixing of diverse swine influenza viruses, leading to the multiple reassortment events associated with the genesis of the S-OIV strain. Domestic pigs have been described as a hypothetical 'mixing-vessel', mediating by reassortment the emergence of new influenza viruses with avian or avian-like genes into the human population, and triggering a pandemic associated with antigenic shift[2]. Previous research has suggested that occupational exposure to pigs increases the risk of swine influenza virus infection, and that swine workers should be considered in any surveillance programs[22].

The emergence of S-OIV provides further evidence of the role of domestic pigs in the ecosystem of influenza A. As reported recently, all three pandemics of the twentieth century seem to have been generated by a series of multiple reassortment events in swine or humans, and to have emerged over a period of years before pandemic recognition[23]. Our results show that the genesis of the S-OIV epidemic followed a similar evolutionary pathway: H1N1 viruses with human pandemic potential had been identified, transmission from swine to humans was known[5] and the disease had been made notifiable. Yet despite widespread influenza surveillance in humans, the lack of systematic swine surveillance allowed for the undetected persistence and evolution of this potentially pandemic strain for many years.

METHODS SUMMARY

We compared 15 newly sequenced Hong Kong swine influenza genomes and two genomes from the S-OIV outbreak with 796 genomes representing the spectrum of influenza A diversity (comprising 285 human, 100 swine and 411 avian isolates). Phylogenetic trees were constructed for each genomic segment independently (Supplementary Fig. 1). Next, for each genomic segment, viruses with known isolation dates that were genetically similar to the current outbreak were identified, and more detailed analysis using a Bayesian 'relaxed molecular clock' approach was performed[24], thereby estimating rates of viral evolution and dates of divergence (Fig. 2). Finally, a similar Bayesian molecular clock approach was applied to the 30 individual viruses isolated from the human outbreak since the end of March 2009 (Supplementary Table 4 and Supplementary Fig. 2). This analysis was performed assuming a model of exponential growth in the number of infections.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 May; accepted 4 June 2009.
Published online 11 June 2009.

1. Centers for Disease Control and Prevention. Swine influenza A (H1N1) infection in two children—Southern California, March–April 2009. *Morb. Mortal. Wkly Rep.* **58,** 400–402 (2009).
2. Shortridge, K. F., Webster, R. G., Butterfield, W. K. & Campbell, C. H. Persistence of Hong Kong influenza virus variants in pigs. *Science* **196,** 1454–1455 (1977).
3. Shope, R. E. & Lewis, P. Swine influenza: experimental transmission and pathology. *J. Exp. Med.* **54,** 349–359 (1931).
4. Brown, I. H., Harris, P. A., McCauley, J. W. & Alexander, D. J. Multiple genetic reassortment of avian and human influenza A viruses in European pigs, resulting in the emergence of an H1N2 virus of novel genotype. *J. Gen. Virol.* **79,** 2947–2955 (1998).
5. Webby, R. J. *et al.* Evolution of swine H3N2 influenza viruses in the United States. *J. Virol.* **74,** 8243–8251 (2000).
6. Newman, A. P. *et al.* Human case of swine influenza A (H1N1) triple reassortant virus infection, Wisconsin. *Emerg. Infect. Dis.* **14,** 1470–1472 (2008).
7. Shinde, V. *et al.* Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. *N. Engl. J. Med.* doi:10.1056/NEJMoa0903812 (in the press).
8. Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. Emergence of a novel swine-origin influenza A (H1N1) virus in humans. *N. Engl. J. Med.* doi:10.1056/NEJMoa0903810 (in the press).
9. Centers for Disease Control and Prevention. Novel influenza A virus infections—2007 case definition. <http://www.cdc.gov/ncphi/disss/nndss/casedef/novel_influenzaA.htm> (24 May 2009).
10. Pensaert, M., Ottis, K., Vanderputte, J., Kaplan, M. M. & Buchmann, P. A. Evidence for the natural transmission of influenza A virus from wild ducks to swine and its potential for man. *Bull. World Health Organ.* **59,** 75–78 (1981).
11. Brown, I. H. The epidemiology and evolution of influenza viruses in pigs. *Vet. Microbiol.* **74,** 29–46 (2000).
12. Peiris, J. S. M. *et al.* Cocirculation of avian H9N2 and contemporary "human" H3N2 influenza A viruses in pigs in southeastern China: potential for genetic reassortment? *J. Virol.* **75,** 9679–9686 (2001).
13. Jung, K. & Song, D. S. Evidence of the cocirculation of influenza H1N1, H1N2 and H3N2 viruses in the pig population of Korea. *Vet. Rec.* **161,** 104–105 (2007).
14. Choi, Y. K. *et al.* H3N2 influenza virus transmission from swine to turkeys, United States. *Emerg. Infect. Dis.* **10,** 2156–2160 (2004).
15. Trifonov, V., Khiabanian, H., Greenbaum, B. & Rabadan, R. The origin of the recent swine influenza A (H1N1) virus infecting humans. *Euro Surveill.* **14,** 19193 (2009).
16. Garten, R. J. *et al.* Antigenic and genetic characteristics of swine-origin 2009 A (H1N1) influenza viruses circulating in humans. *Science* doi:10.1126/science.1176225 (in the press).
17. Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis of high virulence of Hong Kong H5N1 influenza A viruses. *Science* **7,** 1840–1842 (2001).
18. Le, Q. M., Sakai-Tagawa, Y., Ozawa, M., Ito, M. & Kawaoka, Y. Selection of H5N1 influenza virus PB2 during replication in humans. *J. Virol.* **83,** 5278–5281 (2009).
19. Obenauer, J. C. *et al.* Large-scale sequence analysis of avian influenza isolates. *Science* **311,** 1576–1580 (2006).
20. Pybus, O. G. *et al.* Phylogenetic estimation of deleterious mutation load in RNA viruses and its contribution to viral evolution. *Mol. Biol. Evol.* **24,** 845–852 (2007).
21. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science* doi:10.1126/science.1176062 (in the press).
22. Myers, K. P. *et al.* Are swine workers in the United States at increased risk of infection with zoonotic influenza virus? *Clin. Infect. Dis.* **42,** 14–20 (2006).
23. Smith, G. J. D. *et al.* Dating the emergence of pandemic influenza viruses. *Proc. Natl Acad. Sci. USA*. (in the press).
24. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. C. Holmes for comments and encouragement. We acknowledge support from The Royal Society of London (A.R. and O.G.P.), the National Institute of Allergy and Infectious Diseases (NIAID) (G.J.D.S. and M.W.), the Biotechnology and Biological Sciences Research Council (BBSRC) (S.J.L.), and the David and Lucile Packard Foundation (M.W.). A.R. works as a part of the Interdisciplinary Centre for Human and Avian Influenza Research (ICHAIR). This study was supported by the National Institutes of Health (NIAID contract HHSN266200700005C) and the Area of Excellence Scheme of the University Grants Committee (grant AoE/M-12/06) of the Hong Kong SAR Government.

Author Contributions J.B., S.J.L., O.G.P., A.R., G.J.D.S., D.V. and M.W. conceived the study, performed analyses, co-wrote the paper, and all contributed equally to this work. J.S.M.P. co-wrote the paper, Y.G. conceived the study and co-wrote the paper, S.B. and J.R. performed analyses, S.K.M. conducted surveillance, and C.L.C. conducted sequencing. All authors commented on and edited the paper.

Author Information Newly reported sequences have been deposited at GenBank under accession numbers GQ229259–GQ229378. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.R. (a.rambaut@ed.ac.uk) or Y.G. (yguan@hku.hk).

METHODS

Sequence selection for phylogenetic analysis. We downloaded 3,986 complete influenza genomes of any subtype and sampling year (2,490 human, 185 swine and 1,311 avian) from the NCBI Influenza Virus Resource[25] on 29 April 2009. Each sequence set was given a unique ID of the form (ID number)_(Subtype)_(Host)_(isolate name), in which the isolate name is in lower case.

To reduce the number of very similar sequences, we listed all isolates in which the coding region in segment 1 (PB2) was at least one nucleotide different from the others. This left 1,759 human, 166 swine and 1,117 avian complete genome sets. Next we sampled the human, swine and avian sets, selecting one genome set per specific host (as defined in the isolate name, for example, chicken, duck), per specific location (for example, state or province), per year (although isolate name synonyms, for example, duck = dk, hongkong = hk were not accounted for). Two avian and four swine sequence sets were removed owing to bad sequences in one or more segments (for example, frameshifts), leaving 286 human (including S-OIVs), 100 swine and 411 avian sequences in the sampled subset. A further outbreak sequence set (A/Canada-ON/RV1527/2009), and the 15 new swine sequence sets were also added, making a total of 813 complete genome sets for analysis. For the more detailed, temporal analyses, all available S-OIV sequences were used.

The nucleotides in the coding regions of segments 1 (PB2), 2 (PB1), 3 (PA) and 5 (NP) were aligned using ClustalW[26] followed by manual alignment to codon position. The full nucleotide sequences of segments 7 (M1 and M2) and 8 (NS1 and NS2) were also aligned using ClustalW, and the sequences were edited such that all of the codons in first open reading frame (ORF) were followed by the remaining codons in the second ORF (that is, nucleotides were not repeated between the two ORFs). The HA and NA genes (segments 4 and 6) were aligned to codon positions using Muscle[27]. Further H1, H3, N1 and N2 only alignments were also performed.

New swine influenza sequences from Hong Kong. To evaluate the evolutionary history of swine/human influenza A H1N1 viruses, 15 viruses isolated from swine in Hong Kong during 1993 to 2009 were sequenced. Viral RNA was directly extracted from infected allantoic fluid or cell culture using QIAamp viral RNA minikit (Qiagen, Inc.). Complementary DNA was synthesized by reverse transcription reaction, and gene amplification by PCR was performed using specific primers for each gene segment. PCR products were purified with the QIAquick PCR purification kit (Qiagen Inc.) and sequenced by synthetic oligonucleotides. Reactions were performed using Big Dye-Terminator v3.1 Cycle Sequencing Reaction Kit on an ABI PRISM 3730 DNA Analyser (Applied Biosystems) following the manufacturer's instructions. All sequences were assembled and edited with Lasergene version 8.0 (DNASTAR). Full genome sequences of these viruses are available for download at GenBank under accession numbers GQ229259–GQ229378.

Molecular evolution and adaptation. We used the programs SLAC (Single-Likelihood Ancestor Counting)[28] and SNAP (Synonymous Non-synonymous Analysis Program)[29] to compare the mean ratio of non-synonymous changes per non-synonymous site to synonymous changes per synonymous site (d_N/d_S) of outbreak versus non-outbreak sequences. SLAC calculates inferred ancestral sequences for each internal node in a phylogeny using a codon model (and disallowing stop codons), and then counts the synonymous and non-synonymous mutations by comparing each codon to its immediate ancestor. SNAP counts the possible synonymous and non-synonymous codon changes across all pairs of sequences.

In brief, we calculated the effect of the excess of non-synonymous changes in the outbreak data as follows. Assume that S is the number of synonymous sites in a data set, N is the number of non-synonymous sites (typically $\sim 3.5S$ for these data), and ω is the d_N/d_S ratio. If the proportional contribution to the overall rate from synonymous sites is s, then the proportional contribution to the overall rate from non-synonymous sites is equal to $(N/S)(\omega)s$. N, S and ω are all readily estimated from the data. Assuming the same rate of synonymous substitution in both the outbreak and reference data sets, the relative rate expected in the outbreak sequences compared to the reference sequences is thus equal to

$$(s+(N/S)(\omega_{\text{outbreak}})s)/(s+(N/S)(\omega_{\text{reference}})s)$$

Phylogenetic analyses. Phylogenetic trees were inferred using the neighbour-joining distance method, with genetic distances calculated by maximum likelihood under the Hasegawa–Kishino–Yano (HKY) model with gamma-distributed rates among sites (HKY+Γ). Parameters of this model were estimated using maximum likelihood on an initial tree. Temporal phylogenies and rates of evolution were inferred using a relaxed molecular clock model that allows rates to vary among lineages within a Bayesian Markov chain Monte Carlo (MCMC) framework[24]. This was used to sample phylogenies and the dates of divergences between viruses from their joint posterior distribution, in which the sequences are constrained by their known date of sampling. A model comprising a codon-position-specific HKY+Γ substitution model was used. The limited sampling timespan of the S-OIV samples required a simpler model to avoid over-parameterization, so a single HKY+Γ model over all sites was used. For the analyses using Bayesian MCMC sampling, in all cases chain lengths of at least 50 million steps were used with a 10% 'burn-in' removed. Furthermore, at least two independent runs of each were performed and compared to ensure adequate sampling.

25. Bao, Y. *et al.* The influenza virus resource at the national center for biotechnology information. *J. Virol.* **82**, 596–601 (2008).
26. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
27. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
28. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).
29. Korber, B. *HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences* (eds Rodrigo, A. G. & Learn, G. H.) Ch. 4, 55–72 (Kluwer Academic Publishers, 2000).